Additional Resources

This module catalogs several of the resources available for teachers and students using the Collaborative Statistics (col10522) textbook and its derivatives. This module provides links to the complementary teacher's guide, supplemental materials including video lectures and additional problem sets, accessibility information, collection version history and errata, and a list of related works and teachers who have adopted them for their courses.

Additional Resources Currently Available

- Glossary
- View or Download This Textbook Online
- Collaborative Statistics Teacher's Guide
- Supplemental Materials
- Video Lectures
- Version History
- Textbook Adoption and Usage
- Additional Technologies and Notes
- Accessibility and Section 508 Compliance

The following section describes some additional resources for learners and educators. These modules and collections are all available on the Connexions website (http://cnx.org/) and can be viewed online, downloaded, printed, or ordered as appropriate.

Glossary

This module contains the entire glossary for the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The glossary is located at http://cnx.org/content/m16129/latest/.

Below are links to additional resources:

Link to the Statistics Glossary by Dr. Philip Stark, UC Berkeley

http:// statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm

Link to Wikipedia

http:// http://www.wikipedia.org/
(Search on "Glossary of probability and statistics.")

View or Download This Textbook Online
The complete contents of this book are available at no cost on the Connexions website at http://cnx.org/content/col10522/latest/. Anybody can view this content free of charge either as an online e-book or a downloadable PDF file. A low-cost printed version of this textbook is also available here.

Collaborative Statistics Teacher's Guide
A complementary Teacher's Guide for Collaborative statistics is available through Connexions at http://cnx.org/content/col10547/latest/. The Teacher's Guide includes suggestions for presenting concepts found throughout the book as well as recommended homework assignments. A low-cost printed version of this textbook is also available here.

Supplemental Materials
This companion to Collaborative Statistics provides a number of additional resources for use by students and instructors based on the award winning Elementary Statistics Sofia online course, also by textbook authors Barbara Illowsky and Susan Dean. This content is designed to complement the textbook by providing video tutorials, course management materials, and sample problem sets. The Supplemental Materials collection can be found at http://cnx.org/content/col10586/latest/.
Video Lectures

- Video Lecture 1: Sampling and Data
- Video Lecture 2: Descriptive Statistics
- Video Lecture 3: Probability Topics
- Video Lecture 4: Discrete Distributions
- Video Lecture 5: Continuous Random Variables
- Video Lecture 6: The Normal Distribution
- Video Lecture 7: The Central Limit Theorem
- Video Lecture 8: Confidence Intervals
- Video Lecture 9: Hypothesis Testing with a Single Mean
- Video Lecture 10: Hypothesis Testing with Two Means
- Video Lecture 11: The Chi-Square Distribution

- Video Lecture 12: Linear Regression and Correlation

Version History
This module contains a listing of changes, updates, and corrections made to the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The Version History is located at http://cnx.org/content/m17360/latest/.

Textbook Adoption and Usage
This module is designed to track the various derivations of the Collaborative Statistics textbook and its various companion resources, as well as keep track of educators who have adopted various versions for their courses. New adopters are encouraged to provide their contact information and describe how they will use this book for their courses. The goal is to provide a list that will allow educators using this book to collaborate, share ideas, and make suggestions for future development of this text. The Adoption and Usage module is located at http://cnx.org/content/m18261/latest/.

Additional Technologies
In order to provide the most flexible learning resources possible, we invite collaboration from all instructors wishing to create customized versions of this content for use with other technologies. For instance, you may be interested in creating a set of instructions similar to this collection's calculator notes. If you would like to contribute to this collection, please use the contact the authors with any ideas or materials you have created.

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit http://cnx.org/content/m17212/latest/.
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit http://cnx.org/content/m17211/latest/.

Author Acknowledgements

This module contains the author acknowledgements for the Collaborative Statistics textbook/collection.

For this second edition, we appreciate the tremendous feedback from De Anza College colleagues and students, as well as from the dozens of faculty around the world who taught out of the first and preliminary editions. We have updated Collaborative Statistics with contributions from many faculty and students. We especially thank Roberta Bloom, who wrote new problems and additional text.

So many students and colleagues have contributed to the text, both the hard copy and open version. We thank the following people for their contributions to the first and/or second editions.

At De Anza College:
Dr. Inna Grushko (deceased), who wrote the glossary; Diane Mathios, who checked every homework problem in the first edition; Kathy Plum, Lenore Desilets, Charles Klein, Janice Hector, Frank Snow, Dr. Lisa Markus, Dr. Vladimir Logvinenko (deceased), Mo Geraghty, Rupinder Sekhon, Javier Rueda, Carol Olmstead; Also, Dr. Jim Lucas and Valerie Hauber of De Anza's Office of Institutional Research, Mary Jo Kane of Health Services; and the thousands of students who have used this text. Many of the students gave us permission to include their outstanding word problems as homework.

Additional thanks:
Dr. Larry Green of Lake Tahoe Community College, Terrie Teegarden of San Diego Mesa College, Ann Flanigan of Kapiolani Community College, Birgit Aquilonius of West Valley College.

The conversion from a for-profit hard copy text to a free open textbook is the result of many individuals and organizations. We particularly thank Dr. Martha Kanter, Hal Plotkin, Dr. Judy Baker, Dr. Robert Maxfield of Maxfield Foundation, Hewlett Foundation, and Connexions.

Finally, we owe much to Frank, Jeffrey, and Jessica Dean and to Dan, Rachel, Matthew, and Rebecca Illowsky, who encouraged us to continue with our work and who had to hear more than their share of “I’m sorry, I can’t” and “Just a minute, I’m working.”

Student Welcome Letter

Dear Student:

Have you heard others say, “You’re taking statistics? That’s the hardest course I ever took!” They say that, because they probably spent the entire course confused and struggling. They were probably lectured to and never had the chance to experience the subject. You will not have that problem. Let’s find out why.

There is a Chinese Proverb that describes our feelings about the field of statistics:

I HEAR, AND I FORGET

I SEE, AND I REMEMBER

I DO, AND I UNDERSTAND

Statistics is a “do” field. In order to learn it, you must “do” it. We have structured this book so that you will have hands-on experiences. They will enable you to truly understand the concepts instead of merely going through the requirements for the course.

What makes this book different from other texts? First, we have eliminated the drudgery of tedious calculations. You might be using computers or graphing calculators so that you do not need to struggle with algebraic manipulations. Second, this course is taught as a collaborative activity. With others in your class, you will work toward the common goal of learning this material.

Here are some hints for success in your class:

- Work hard and work every night.
- Form a study group and learn together.
- Don’t get discouraged - you can do it!
- As you solve problems, ask yourself, “Does this answer make sense?”
- Many statistics words have the same meaning as in everyday English.

- Go to your teacher for help as soon as you need it.
- Don’t get behind.
- Read the newspaper and ask yourself, “Does this article make sense?”
- Draw pictures - they truly help!

Good luck and don’t give up!

Sincerely,
Susan Dean and Barbara Illowsky

De Anza College
21250 Stevens Creek Blvd.
Cupertino, California 95014

Introduction
This module provides a brief introduction to the field of statistics, including examples of how these topics shows up in a variety of real-life examples.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

Introduction

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data are.

Statistics

This module introduces the concept of statistics, specifically the ability to use statistics to describe data (descriptive statistics) as well as draw conclusions (inferential statistics). An optional classroom exercise is included.

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

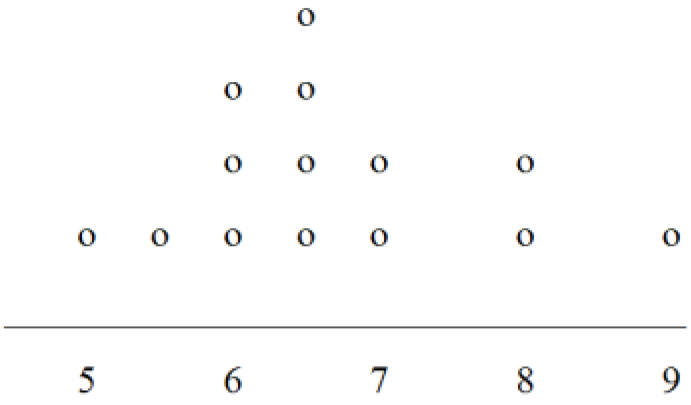
Optional Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:

Frequency of Average Time (in Hours) Spent Sleeping per Night



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that the conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Levels of Measurement and Statistical Operations

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- Nominal scale level
- Ordinal scale level
- Interval scale level
- Ratio scale level

Data that is measured using a **nominal scale** is qualitative. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory” and “unsatisfactory.” These responses are ordered from the most desired response by the cruise lines to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40 degrees is equal to 100 degrees minus 60 degrees. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations but one type of comparison cannot be done. Eighty degrees C is not 4 times as hot as 20° C (nor is 80° F 4 times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or 4 to 1).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data but, in addition, it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams were machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points.

Ratios can be calculated. The smallest score for ratio data is 0. So 80 is 4 times 20. The score of 80 is 4 times better than the score of 20.

Exercises

What type of measure scale is being used? Nominal, Ordinal, Interval or Ratio.

1. High school men soccer players classified by their athletic ability: Superior, Average, Above average.
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box.
4. Social security numbers.
5. Incomes measured in dollars
6. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied.
7. Political outlook: extreme left, left-of-center, right-of-center, extreme right.
8. Time of day on an analog watch.
9. The distance in miles to the closest grocery store.
10. The dates 1066, 1492, 1644, 1947, 1944.

11. The heights of 21 – 65 year-old women.
12. Common letter grades A, B, C, D, F.

Answers 1. ordinal, 2. interval, 3. nominal, 4. nominal, 5. ratio, 6. ordinal, 7. nominal, 8. interval, 9. ratio, 10. interval, 11. ratio, 12. ordinal

Glossary

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Statistic

A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

Probability
This module introduces the concept of probability as a mathematical measure of randomness, including a number of real-world applications.

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin 4 times, the outcomes may not be 2 heads and 2 tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Glossary

Probability

A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930’s): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1;$.
- If A and B are any two mutually exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$.

- $P(S) = 1$.

Key Terms

This module introduces a number of key terms related to statistical sampling and data.

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that is a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters like X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:
Mean and Average
The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:

Exercise:

Problem:

Define the key terms from the following study: We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Optional Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Proportion

- As a number: A proportion is the number of successes divided by the total number in the sample.
- As a probability distribution: Given a binomial random variable (RV), $X \sim B(n, p)$, consider the ratio of the number X of successes in n Bernouli trials to the number n of trials. $P\prime = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, P’ $\sim N\left(p, \frac{\text{pq}}{n}\right)$.

Data

This module introduces the concepts of qualitative data, quantitative continuous data, and quantitative discrete data as used in statistics. Sample problems are included.

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in the numbers $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Note: In this course, the data used is mainly quantitative. It is easy to calculate statistics (like the mean or proportion) from numbers. In the chapter **Descriptive Statistics**, you will be introduced to stem plots, histograms and box plots all of which display quantitative data. Qualitative data is discussed at the end of this section through graphs.

Example:
Data Sample of Quantitative Discrete Data
The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example:
Data Sample of Quantitative Continuous Data
The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example:
Data Sample of Qualitative Data
The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:
Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.
10. Weight of sumo wrestlers.
11. Amount of money won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Solution:

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

Qualitative Data Discussion

Below are tables of part-time vs full-time students at De Anza College in Cupertino, CA and Foothill College in Los Altos, CA for the Spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

De Anza College

	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%

Total	14,183	100%

Foothill College

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning what graphs to use. Below are pie charts and bar graphs, two graphs that are used to display qualitative data.

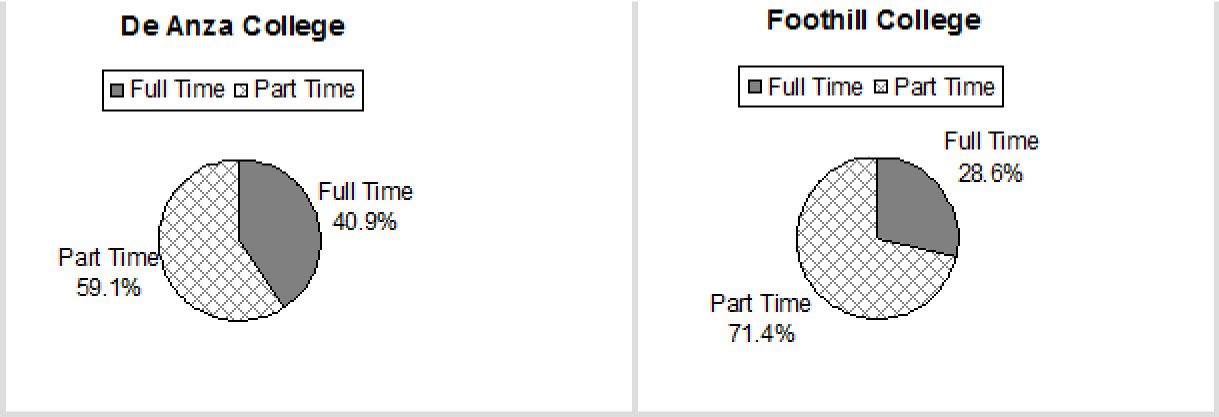
In a **pie chart**, categories of data are represented by wedges in the circle and are proportional in size to the percent of individuals in each category.

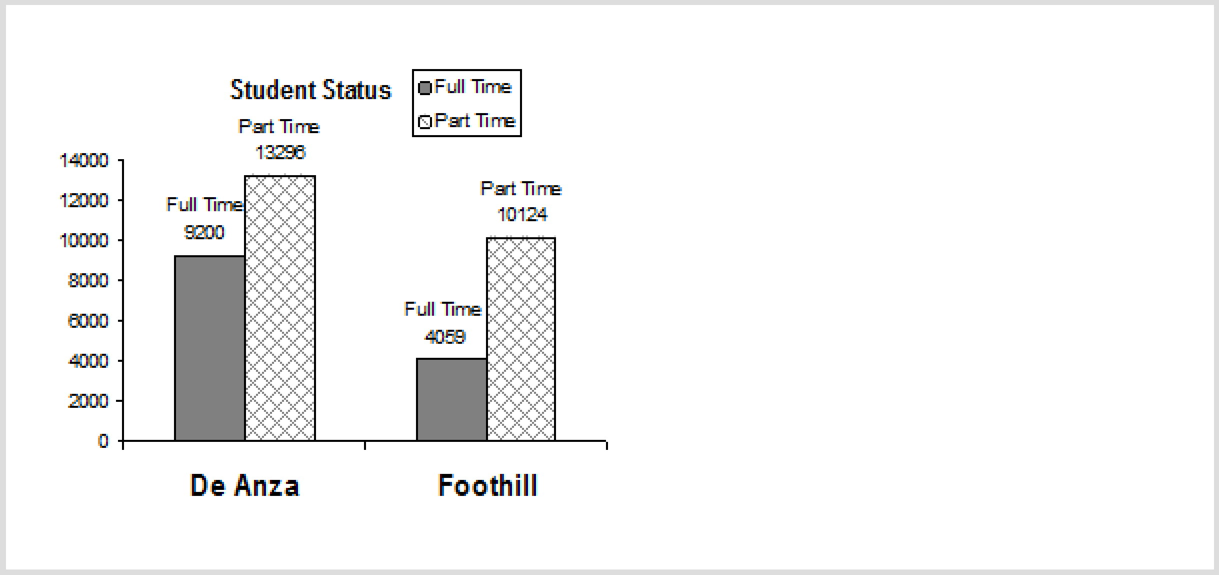
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at the graphs and determine which graph (pie or bar) you think displays the comparisons better. This is a matter of preference.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.



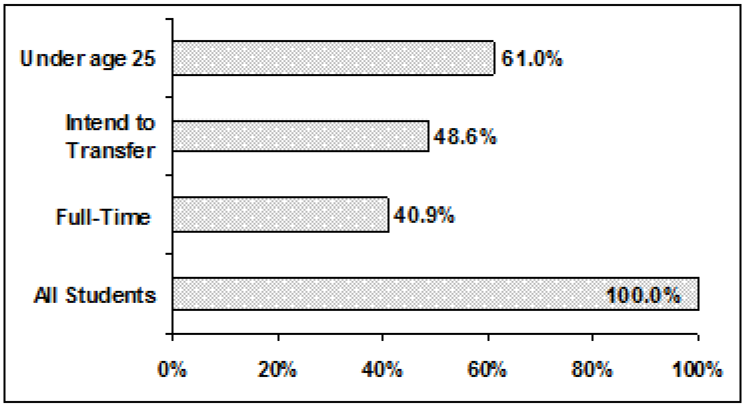


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

De Anza College Spring 2010

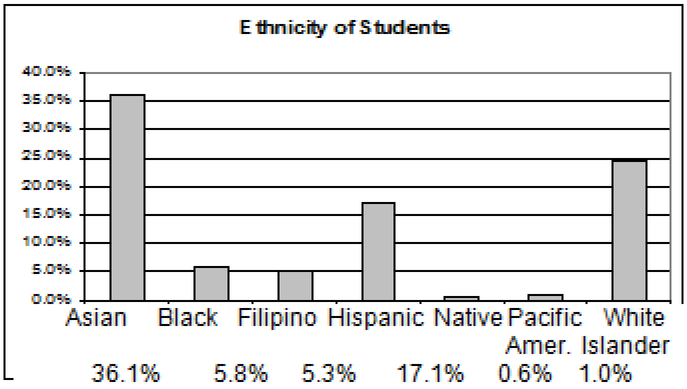


Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. Create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

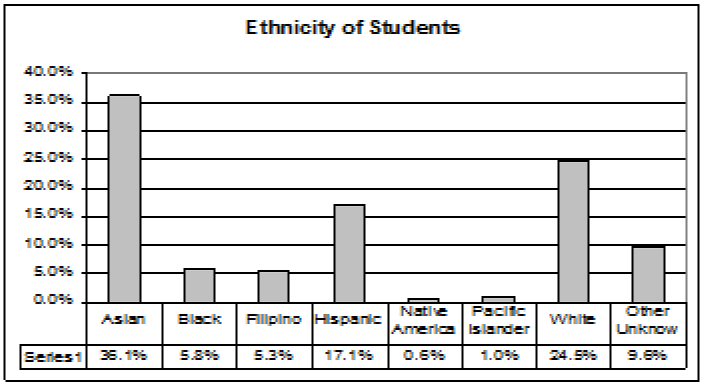
Missing Data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)



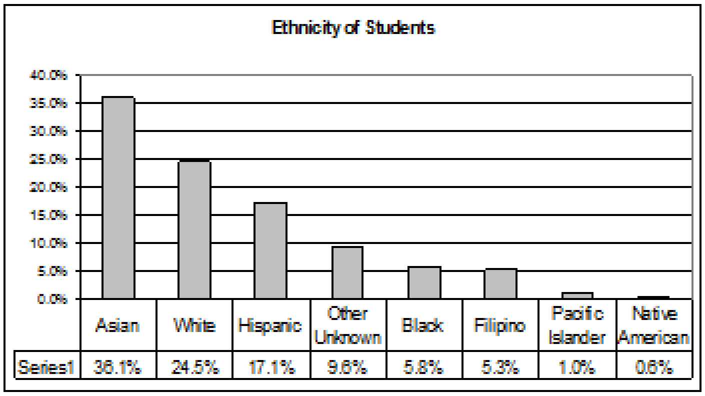
Bar graph Without Other/Unknown Category

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been added back in. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0% particularly). This is important to know when we think about what the data are telling us.

This particular bar graph can be hard to understand visually. The graph below it is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.



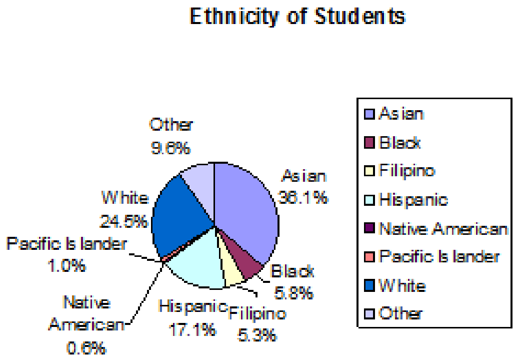
Bar Graph With Other/Unknown Category

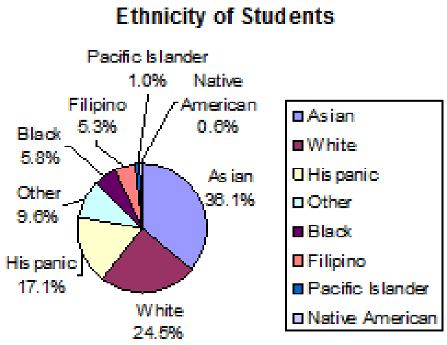


Pareto Chart With Bars Sorted By Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category added back in (since the percentages must add to 100%). The chart on the right is organized having the wedges by size and makes for a more visually informative graph than the unsorted, alphabetical graph on the left.





Glossary

Continuous Random Variable

A random variable (RV) whose outcomes are measured.

Example:
The height of trees in the forest is a continuous RV.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Discrete Random Variable

A random variable (RV) whose outcomes are counted.

Qualitative Data

See **Data**.

Quantitative Data

See **Data**.

Sampling

This module introduces the concept of statistical sampling. Students are taught the difference between a simple random sample, stratified sample, cluster sample, systematic sample, and convenience sample. Example problems are provided, including an optional classroom activity.

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

ID	Name
00	Anselmo

ID	Name
01	Bautista
02	Bayani
03	Cheng
04	Cuarismo
05	Cuningham
06	Fontecha
07	Hong
08	Hoobler
09	Jiao
10	Khan
11	King
12	Legeny
13	Lundquist
14	Macierz
15	Motogawa
16	Okimoto
17	Patel

ID	Name
18	Price
19	Quizon
20	Reyes
21	Roquero
22	Roth
23	Rowell
24	Salangsang
25	Slade
26	Stracher
27	Tallai
28	Tran
29	Wai
30	Wood

Class Roster

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360 .99832 .14669 .51470 .40581 .73381 .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. For example, divide your college faculty by department. The departments are the clusters. Number each department and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is nonrandom is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favors certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (for example, they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once using with replacement is very low.

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To 4 decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To 4 decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example:
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

1. stratified
2. cluster
3. stratified
4. systematic
5. simple random
6. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will seem natural.

Example:

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:

$128 $87 $173 $116 $130 $204 $147 $189 $93 $153

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:
$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

Exercise:

Problem:

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of

part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he/she has a corresponding number. The students spend:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Exercise:

Problem: Is the sample biased?

Solution:

The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation
This module discusses statistical variability within data and samples. Students will be given the opportunity to see this variability in action through participation in an optional classroom exercise. This module also has a section that discusses Critical Evaluation.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population are different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariable biased because people choose to respond or not.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice.** Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to critically evaluate the statistical studies we read about and analyze before accepting the results of the study. Common problems to be aware of include

- Problems with Samples: A sample should be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-Selected Samples: Responses only by people who choose to respond, such as call-in surveys are often unreliable.
- Sample Size Issues: Samples that are too small may be unreliable. Larger samples are better if possible. In some situations, small samples are unavoidable and can still be used to draw conclusions, even though larger samples are better. Examples: Crash testing cars, medical testing for rare conditions.
- Undue influence: Collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship through a different variable.
- Self-Funded or Self-Interest Studies: A study performed by a person or organization in order to support their claim. Is the study impartial?

Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- Misleading Use of Data: Improperly displayed graphs, incomplete data, lack of context.
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Glossary

Population
: The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Sample
: A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Answers and Rounding Off

This module briefly explains the correct way to round off answers when working with statistical data.

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round only the final answer. Do not round any intermediate results, if possible. If it becomes necessary to round intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores 4, 6, 9 is 6.3, rounded to the nearest tenth, because the data are whole numbers. Most answers will be rounded in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Frequency

This module introduces the concepts of frequency, relative frequency, and cumulative relative frequency, and the relationship between these measures. Students will have the opportunity to interpret data through the sample problems provided.

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5 6 3 3 2 4 7 5 2 3 5 6 5 4 4 3 5 2 5 3

Below is a frequency table listing the different data values in ascending order and their frequencies.

DATA VALUE	**FREQUENCY**
2	3
3	5
4	3
5	6
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the fraction or proportion of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours w/ Relative Frequency

The sum of the relative frequency column is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

DATA VALUE	**FREQUENCY**	**RELATIVE FREQUENCY**	**CUMULATIVE RELATIVE FREQUENCY**
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note: Because of rounding, the relative frequency column may not always sum to one and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	**FREQUENCY**	**RELATIVE FREQUENCY**	**CUMULATIVE RELATIVE FREQUENCY**
59.95 - 61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95 - 63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08
63.95 - 65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95 - 67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
67.95 - 69.95	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
69.95 - 71.95	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
71.95 - 73.95	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99
73.95 - 75.95	1	$\frac{1}{100}$ = 0.01	0.99 + 0.01 = 1.00
	Total = 100	Total = 1.00	

Frequency Table of Soccer Player Height

The data in this table has been **grouped** into the following intervals:

- 59.95 - 61.95 inches
- 61.95 - 63.95 inches
- 63.95 - 65.95 inches
- 65.95 - 67.95 inches
- 67.95 - 69.95 inches

- 69.95 - 71.95 inches
- 71.95 - 73.95 inches
- 73.95 - 75.95 inches

Note: This example is used again in the Descriptive Statistics chapter, where the method used to compute the intervals will be explained.

In this sample, there are **5** players whose heights are between 59.95 - 61.95 inches, **3** players whose heights fall within the interval 61.95 - 63.95 inches, **15** players whose heights fall within the interval 63.95 - 65.95 inches, **40** players whose heights fall within the interval 65.95 - 67.95 inches, **17** players whose heights fall within the interval 67.95 - 69.95 inches, **12** players whose heights fall within the interval 69.95 - 71.95, 7 players whose height falls within the interval 71.95 - 73.95, and **1** player whose height falls within the interval 73.95 - 75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:

Exercise:

Problem:

From the table, find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 males whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Example:

Exercise:

Problem:

From the table, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Example:
Exercise:

Problem:

Use the table of heights of the 100 male semiprofessional soccer players. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is:
2. The percentage of heights that are from 67.95 to 73.95 inches is:
3. The percentage of heights that are more than 65.95 inches is:
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution:

1. 29%
2. 36%
3. 77%

4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Optional Collaborative Classroom Exercise

Exercise:

Problem:

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class has 0 siblings?
2. What percentage of the students has from 1 to 3 siblings?
3. What percentage of the students has fewer than 3 siblings?

Example:

Nineteen people were asked how many miles, to the nearest mile they commute to work each day. The data are as follows:
2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10
The following table was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Frequency of Commuting Distances

Exercise:

Problem:

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute 3 miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute 5 or 7 miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (does not include 5 and 13 miles)?

Solution:

1. No. Frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. Frequency for 3 miles should be 1; for 2 miles (left out), 2. Cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.
3. $\frac{5}{19}$
4. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Glossary

Frequency
: The number of times a value of the data occurs.

Relative Frequency
: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Cumulative Relative Frequency
: The term applies to an ordered set of observations from smallest to largest. The Cumulative Relative Frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Summary

This module provides an outline/review of key concepts related to statistical sampling and data.

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
  - Discrete (You count it.)
  - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement**: A member of the population may be chosen more than once
- **Without Replacement**: A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
  - Simple random sample
  - Stratified sample
  - Cluster sample
  - Systematic sample
- Not Random
  - Convenience sample

Frequency (freq. or f)

- The number of times an answer occurs

Relative Frequency (rel. freq. or RF)

- The proportion of times an answer occurs
- Can be interpreted as a fraction, decimal, or percent

Cumulative Relative Frequencies (cum. rel. freq. or cum RF)

- An accumulation of the previous relative frequencies

Practice: Sampling and Data
This module provides an opportunity for students to practice concepts related to statistical sampling and data. Given a sample data set, the student will practice constructing frequency tables, differentiating between key terms, and comparing sampling techniques.

Student Learning Outcomes

- The student will construct frequency tables.
- The student will differentiate between key terms.
- The student will compare sampling techniques.

Given

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average(mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Organize the Data

Complete the tables below using the data provided.

Survival Length (in months)	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher A

Survival Length (in months)	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher B

Key Terms

Define the key terms based upon the above example for Researcher A.

Exercise:

Problem: Population

Exercise:

Problem: Sample

Exercise:

Problem: Parameter

Exercise:

Problem: Statistic

Exercise:

Problem: Variable

Exercise:

Problem: Data

Discussion Questions

Discuss the following questions and then answer in complete sentences.

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect? Why?

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem: How could the researchers gather random data?

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Homework

This module presents students with a number of problems related to statistical sampling and data. In particular, students are asked to demonstrate understanding of concepts such as frequency, relative frequency, and cumulative relative frequency, random samples, quantitative vs. qualitative data, continuous vs. discrete data, and other key terms related to sampling and data.

Exercise:

Problem: For each item below:

- **i**Identify the type of data (quantitative - discrete, quantitative - continuous, or qualitative) that would be used to describe a response.
- **ii**Give an example of the data.

- **a**Number of tickets sold to a concert
- **b**Amount of body fat
- **c**Favorite baseball team
- **d**Time in line to buy groceries
- **e**Number of students enrolled at Evergreen Valley College
- **f**Most–watched television show
- **g**Brand of toothpaste
- **h**Distance to the closest movie theatre
- **i**Age of executives in Fortune 500 companies
- **j**Number of competing computer spreadsheet software packages

---

Solution:

- **a**quantitative - discrete
- **b**quantitative - continuous
- **c**qualitative
- **d**quantitative - continuous
- **e**quantitative - discrete
- **f**qualitative
- **g**qualitative

- **h**quantitative - continuous
- **i**quantitative - continuous
- **j**quantitative - discrete

Exercise:

Problem:

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Part-time Student Course Loads

- **a**Fill in the blanks in the table above.
- **b**What percent of students take exactly two courses?
- **c**What percent of students take one or two courses?

Exercise:

Problem:

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnoses. The (incomplete) results are shown below:

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

- **a**Fill in the blanks in the table above.
- **b**What percent of adults flossed six times per week?
- **c**What percent flossed at most three times per week?

Solution:

- **a** Cum. Rel. Freq. for 0 is 0.4500
  Rel. Freq. for 1 is 0.3000 and Cum. Rel. Freq. for 1 or less is 0.7500

Freq. for 3 is 11 and Rel. Freq. is 0.1833
Cum. Rel. Freq. for 6 or less is 0.9833
Cum. Rel. Freq. for 7 or less is 1

- **b**5.00%
- **c**93.33%

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to optimally plan their ski classes. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**Children who take ski or snowboard lessons
- **b**A group of these children
- **c**The population mean
- **d**The sample mean
- **e**X = the age of one child who takes the first ski or snowboard lesson
- **f**Values for X, such as 3, 7, etc.

Exercise:

Problem:

A cardiologist is interested in the mean recovery period for her patients who have had heart attacks. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year for their clients, so that they can determine the costs of health insurance. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic

- **e**Variable
- **f**Data

Solution:

- **a**The clients of the insurance companies
- **b**A group of the clients
- **c**The mean health costs of the clients
- **d**The mean health costs of the sample
- **e**X = the health costs of one client
- **f**Values for X, such as 34, 9, 82, etc.

Exercise:

Problem:

A politician is interested in the proportion of voters in his district that think he is doing a good job. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marriage counselor is interested in the proportion the clients she counsels that stay married. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample

- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**All the clients of the counselor
- **b**A group of the clients
- **c**The proportion of all her clients who stay married
- **d**The proportion of the sample who stay married
- **e**X = the number of couples who stay married
- **f**yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people that will vote for a particular cause. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marketing company is interested in the proportion of people that will buy a particular product. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**All people (maybe in a certain geographic area, such as the United States)
- **b**A group of the people
- **c**The proportion of all people who will buy the product
- **d**The proportion of the sample who will buy the product
- **e**X = the number of people who will buy it
- **f**buy, not buy

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys 6 flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

- **a**Using complete sentences, list three things wrong with the way the survey was conducted.
- **b**Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Exercise:

Problem:

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in 3 – 5 complete sentences. Make the description detailed.

Exercise:

Problem:

Suppose you want to determine the mean number of cans of soda drunk each month by persons in their twenties. Describe a possible sampling method in 3 - 5 complete sentences. Make the description detailed.

Exercise:

Problem:

771 distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in the table below. (Source: http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus).

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%

Took DL courses to fulfill transfer requirements	71%

LBCC Distance Learning Survey Results

- **a**What percent of the students surveyed do not have a computer at home?
- **b**About how many students in the survey live at least 16 miles from campus?
- **c**If the same survey was done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Solution:

- **a**4%
- **b**100

Exercise:

Problem:

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10

The following table was produced:

Data	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0	2	$\frac{2}{19}$	0.1053

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Frequency of Immigrant Survey Responses

- **a**Fix the errors on the table. Also, explain how someone might have arrived at the incorrect number(s).
- **b**Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
- **c**Fix the statement above to make it correct.
- **d**What fraction of the people surveyed have lived in the U.S. 5 or 7 years?
- **e**What fraction of the people surveyed have lived in the U.S. at most 12 years?
- **f**What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
- **g**What fraction of the people surveyed have lived in the U.S. from 5 to 20 years, inclusive?

Exercise:

Problem:

A “random survey” was conducted of 3274 people of the “microprocessor generation” (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users. (*Source: San Jose Mercury News*)

- **a**Do you consider the sample size large enough for a study of this type? Why or why not?
- **b**Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

Additional information: The survey was reported by Intel Corporation of individuals who visited the Los Angeles Convention Center to see the Smithsonian Institure's road show called “America’s Smithsonian.”

- **c**With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
- **d**With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise:

Problem:

- **a**List some practical difficulties involved in getting accurate results from a telephone survey.
- **b**List some practical difficulties involved in getting accurate results from a mailed survey.
- **c**With your classmates, brainstorm some ways to overcome these problems if you needed to conduct a phone or mail survey.

Try these multiple choice questions

The next four questions refer to the following: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

- **A**All Lake Tahoe Community College students
- **B**All Lake Tahoe Community College English students
- **C**All Lake Tahoe Community College students in her classes
- **D**All Lake Tahoe Community College math students

Solution:

D

Exercise:

Problem: Consider the following:

X **= number of days a Lake Tahoe Community College math student is absent**

In this case, X is an example of a:

- **A**Variable
- **B**Population
- **C**Statistic
- **D**Data

Solution:

A

Exercise:

Problem:

The instructor takes her sample by gathering data on 5 randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

- **A**Cluster sampling
- **B**Stratified sampling
- **C**Simple random sampling
- **D**Convenience sampling

Solution:

B

Exercise:

Problem:

The instructor’s sample produces an mean number of days absent of 3.5 days. This value is an example of a

- **A**Parameter
- **B**Data
- **C**Statistic
- **D**Variable

Solution:

C

The next two questions refer to the following relative frequency table on hurricanes that have made direct hits on the U.S between 1851 and 2004. Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. (http://www.nhc.noaa.gov/gifs/table5.gif)

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Frequency of Hurricane Direct Hits

Exercise:

Problem:

What is the relative frequency of direct hits that were category 4 hurricanes?

- **A**0.0768
- **B**0.0659
- **C**0.2601
- **D**Not enough information to calculate

Solution:

B

Exercise:

Problem:

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

- **A**0.3480
- **B**0.9231
- **C**0.2601
- **D**0.3370

Solution:

B

The next three questions refer to the following: A study was done to determine the age, number of times per week and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise:

Problem: "‘Number of times per week’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

B

Exercise:

Problem: The sampling method was:

- **A**simple random
- **B**systematic
- **C**stratified
- **D**cluster

Solution:

B

Exercise:

Problem: "‘Duration (amount of time)’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

C

Exercises 28 and 29 are not multiple choice exercises.

Exercise:

Problem:

Name the sampling method used in each of the following situations:

- **A**A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport’s service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers sitting near gates and who are not taking naps while they wait.

- **B**A teacher wants to know if her students are doing homework so she randomly selects rows 2 and 5, and then calls on all students in row 2 and all students in row 5 to present the solution to homework problems to the class.
- **C**The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out which asks for information about age, as well as about other variables of interest.
- **D**The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether the books are checked out by an adult or a child. She records this data for every 4th patron who checks out books.
- **E**A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked who he/she intends to vote for and whether the debate changed his/her opinion of the candidates.

** Contributed by Roberta Bloom

Solution:

- **A**Convenience
- **B**Cluster
- **C**Stratified
- **D**Systematic
- **E**Simple Random

Exercise:

Problem:

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following 7 subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these 7 textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

** Contributed by Roberta Bloom

Solution:

The answer below contains some of the issues that students might discuss for this problem. Individual student's answers may also identify other issues that pertain to this problem that are not included in the answer below.

The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled 7 subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The 7 subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and many other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also only looked at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed and more students nationwide selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the popular and less popular textbooks.

Introduction

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics"**. You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

Displaying Data

This module provides a brief introduction into the ways graphs and charts can be used to provide visual representations of data.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar chart, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), pie charts, and the boxplot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

Graphing Qualitative Variables

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple?s market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market, and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owners, a previous Windows owner, or a new computer purchaser. This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous iMac users. This situation may be contrasted with quantitative data, such as a person?s weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. [link] shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" of $0.17 = 85/500$.

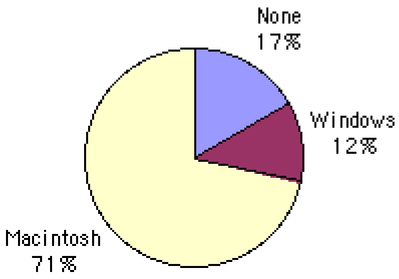
Previous Ownership	Frequency	Relative Frequency

Previous Ownership	**Frequency**	**Relative Frequency**
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Frequency Table for the Mac Data

Pie Charts

The pie chart in [link] shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.



Pie chart of iMac purchases illustrating frequencies of

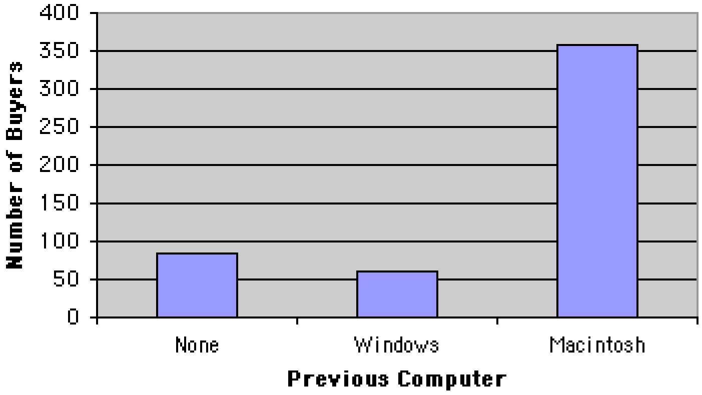
previous computer
ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted ""The only worse design than a pie chart is several of them"".

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have accord since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar Charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in [link]. Frequencies are shown on the Y axis and the type of computer previously owned is shown on the X axis. Typically the Y-axis shows the number of observations rather than the percentage of observations in each category as is typical in pie charts.

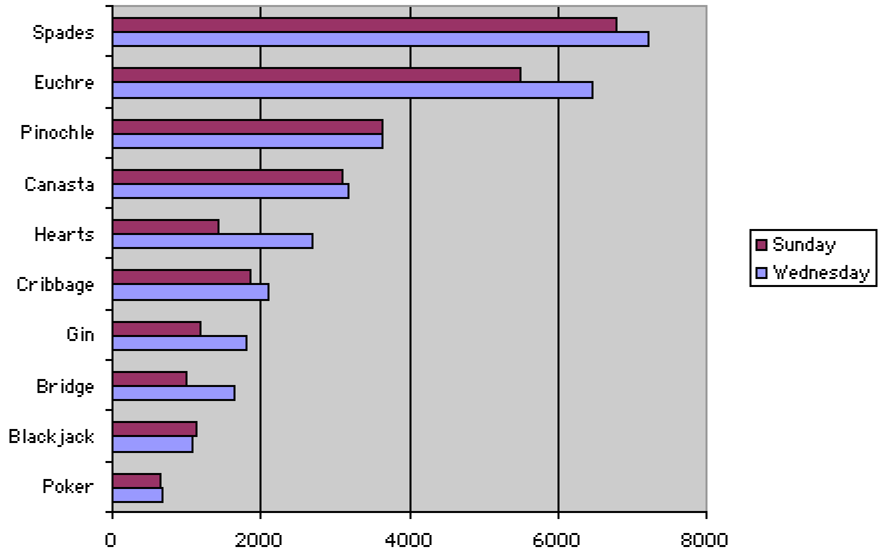


Bar chart of iMac purchases as a function of previous computer ownership.

Note:EIA guidelines for statistical graphs.

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the **distributions** of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. [link] shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday on a day in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.



A bar chart of the number of people playing different card games on Sunday and Wednesday.

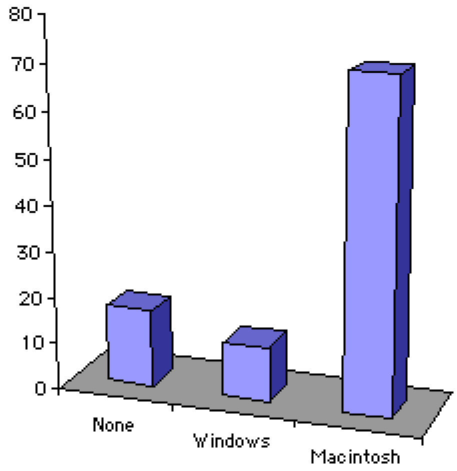
The bars in [link] are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. Such **horizontal** bar charts may be contrasted with **vertical** bar charts like the one in [link].

Note: Another example using bar charts to compare distributions.

We'll have more to say about bar charts when we consider numerical quantities later in this chapter. (See Bar Charts.)

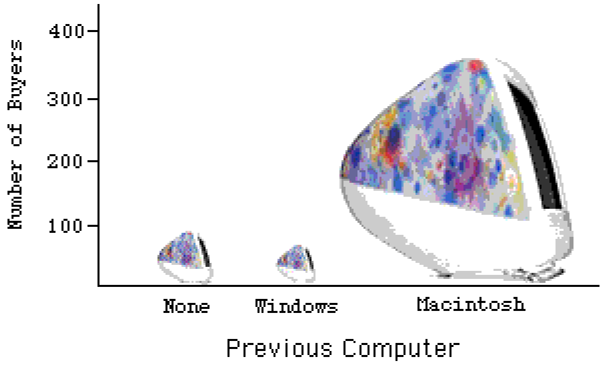
Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts like the one shown in [link] are usually not as effective as their two-dimensional counterparts.



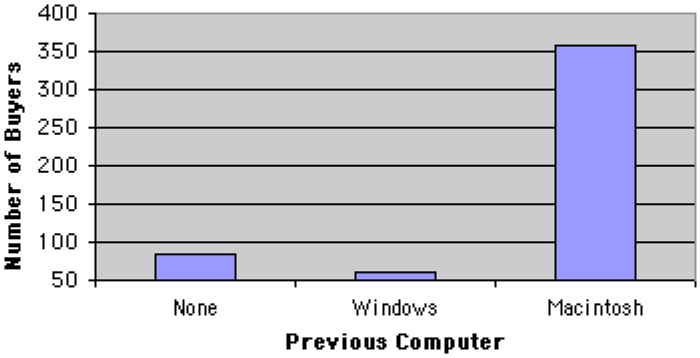
A three-dimensional version of [link].

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, [link] presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet [link] is misleading because the viewer's attention will be captured by areas. This can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in [link] is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use [link] instead of [link]! Edward Tufte coined the term **lie factor** to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.



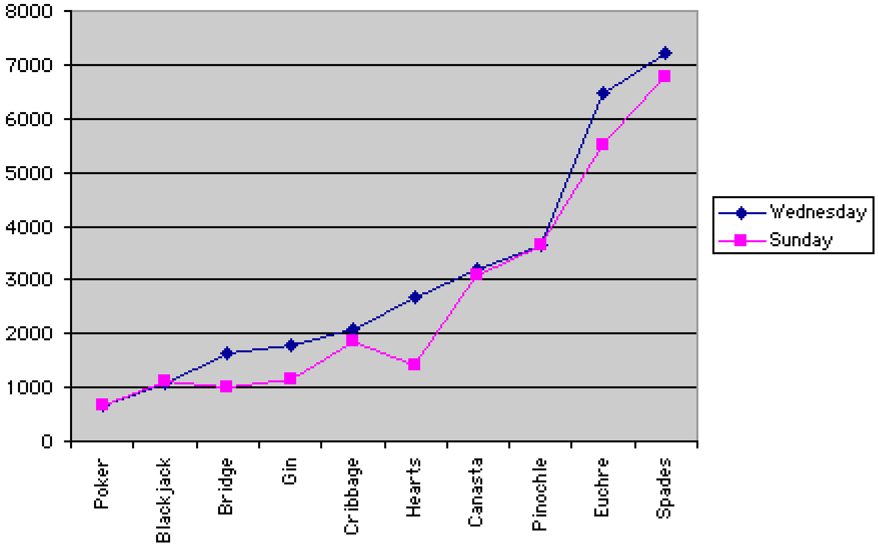
A redrawing of [link] with a lie factor greater than 8.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, this number should be zero. [link] shows the iMac data with a baseline of 50. Once again, the difference in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.



A redrawing of [link] with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). [link] inappropriately shows a line graph of the card game data from Yahoo. The drawback to [link] is that it gives the false impression that the games are naturally ordered in a numerical way.



A line graph of the number of people playing different card games on Sunday and Wednesday.

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

Applet failed to run. No Java plug-in was found.

Histograms
This module provides an overview of Descriptive Statistics: Histogram as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **Frequency** or **relative frequency**. The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on Sampling and Data, we defined frequency as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
Equation:

$$\text{RF} = \frac{f}{n}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$f = 3$, $n = 40$, and $\text{RF} = \frac{f}{n} = \frac{3}{40} = 0.075$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100 % are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 - 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 - 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 - .0005 = 0.9995). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 (2 - 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example:

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.

60 60.5 61 61 61.5

63.5 63.5 63.5

64 64 64 64 64 64 64 64.5 64.5 64.5 64.5 64.5 64.5 64.5 64.5

66 66 66 66 66 66 66 66 66 66 66.5 66.5 66.5 66.5 66.5 66.5 66.5 66.5

66.5 66.5 66.5 67 67 67 67 67 67 67 67 67 67 67 67 67.5 67.5 67.5 67.5

67.5 67.5 67.5

68 68 69 69 69 69 69 69 69 69 69 69 69.5 69.5 69.5 69.5 69.5

70 70 70 70 70 70 70.5 70.5 70.5 71 71 71

72 72 72 72.5 72.5 73 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 - 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74. 74+ 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.

Equation:

$$\frac{74.05-59.95}{8}=1.76$$

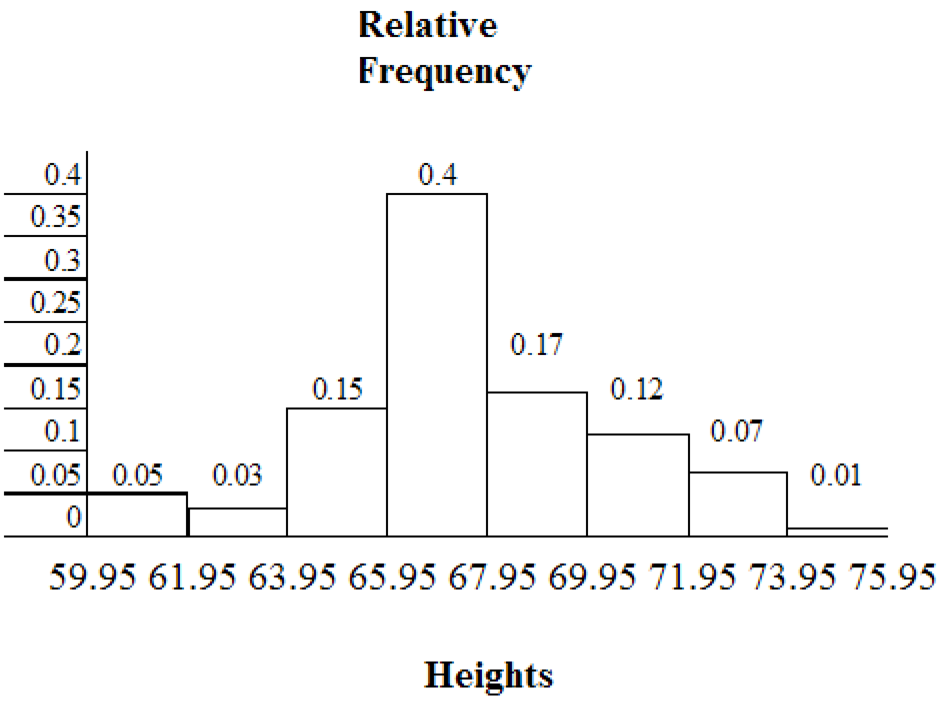
Note: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95

- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95 - 61.95. The heights that are 63.5 are in the interval 61.95 - 63.95. The heights that are 64 through 64.5 are in the interval 63.95 - 65.95. The heights 66 through 67.5 are in the interval 65.95 - 67.95. The heights 68 through 69.5 are in the interval 67.95 - 69.95. The heights 70 through 71 are in the interval 69.95 - 71.95. The heights 72 through 73.5 are in the interval 71.95 - 73.95. The height 74 is in the interval 73.95 - 75.95.
The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Example:
The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books

are counted.
1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4
5 5 5 5 5
6 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _______ to _______, the 5 in the middle of the interval from _______ to _______, and the _______ in the middle of the interval from _______ to _______ .

Solution:

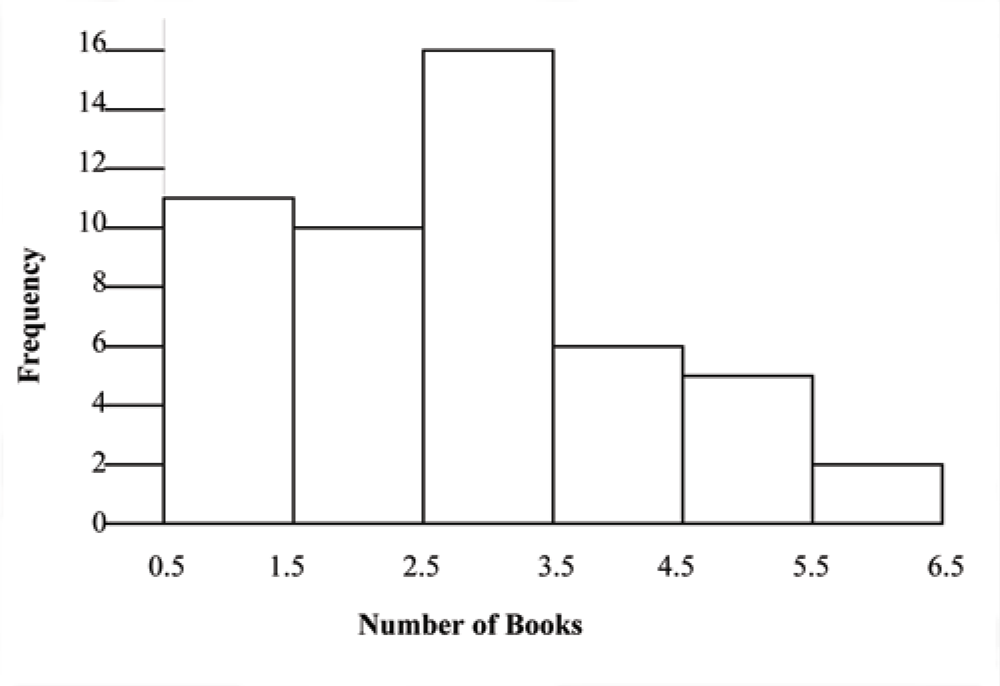
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

Equation:

$$\frac{6.5 - 0.5}{\text{bars}} = 1$$

where 1 is the width of a bar. Therefore, bars $= 6$.
The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



Using the TI-83, 83+, 84, 84+ Calculator Instructions
Go to the Appendix (14:Appendix) in the menu on the left. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.

- Press Y=. Press CLEAR to clear out any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6
- Into L2, enter 11, 10, 16, 6, 5, 2
- Press WINDOW. Make Xmin = .5, Xmax = 6.5, Xscl = (6.5 - .5)/6, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1

- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH
- Use the TRACE key and the arrow keys to examine the histogram.

Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

Glossary

Frequency

The number of times a value of the data occurs.

Relative Frequency

The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

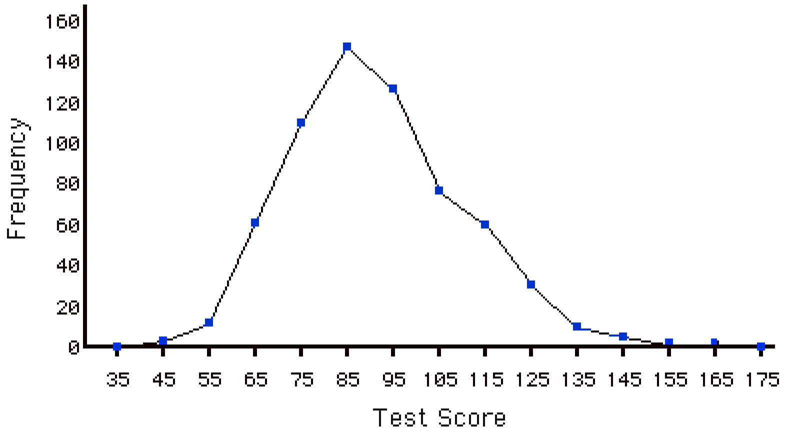
Frequency Polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful in comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

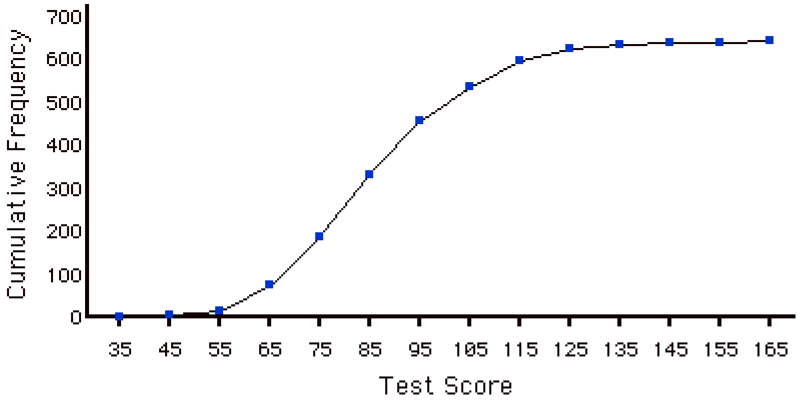
A frequency polygon for 642 psychology test scores is shown in [link]. The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 150 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from [link]. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is **skewed**.



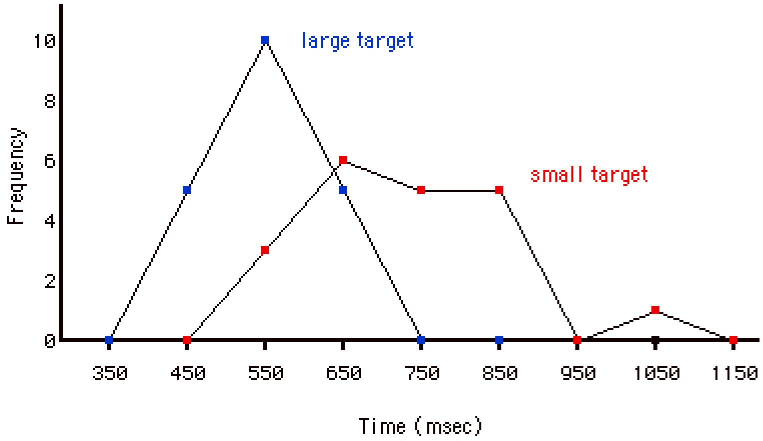
Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in [link]. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled "35," three in the interval "45,"and 10 in the interval "55."Therefore the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.



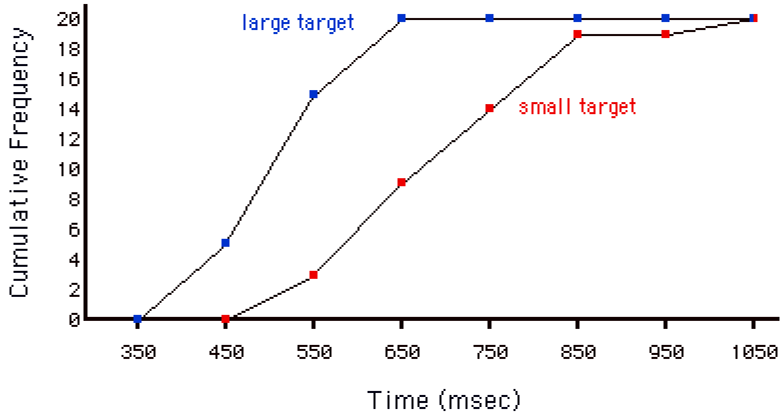
Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. [link] provides an example. The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in [link]. The figure shows that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.



Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in [link] using the same data from the mouse task. The difference in distributions for the two targets is again evident.



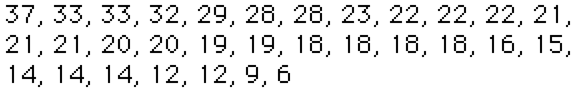
Overlaid cumulative frequency polygons.

Stem and Leaf Displays

A **stem and leaf** display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, an example will get us started. Consider [link]. It shows the number of touchdown (TD) passes [footnote] thrown by each of the 31 teams in the National Football League in the 2000 season.

Touchdown Pass: In American football, a touchdown pass occurs when a completed pass results in a touchdown. The pass may be to a player in the end zone or to a player who subsequently runs into the end zone. A touchdown is worth 6 points and allows for a chance at one (and by some rules two) additional point(s).



37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21,
21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6

Number of touchdown passes.

A stem and leaf display of the data is shown in the [link] below. The left portion of the table contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits. A stem of 3 (for example) can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

3|2337
2|001112223889

1|2244456888899
0|69

Stem and leaf display showing the number of passing touchdowns.

To make this clear, let us examine this [link] more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD passes for the first four teams in the table. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.).

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in [link] than in the [link]. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. The [link] below shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in the [link]. The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

<table>
<tr><td>3|7</td></tr>
<tr><td>3|233</td></tr>
<tr><td>2|889</td></tr>
<tr><td>2|001112223</td></tr>
<tr><td>1|56888899</td></tr>
<tr><td>1|22444</td></tr>
<tr><td>0|69</td></tr>
</table>

Stem and leaf display with the stems split in two.

The [link] with stem and leaf split in two is more revealing than the simpler [link] before because the simpler table lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a **back to back stem and leaf graph**. The [link] below shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with 11, 12, and 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

1998		2000
11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

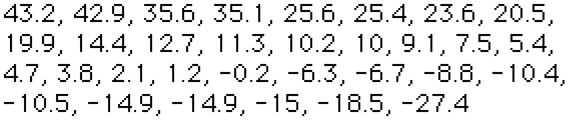
Back to back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

This [link] helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

[link] shows data from a study on aggressive thinking. Each value is the mean difference over a series of trials between the time it took an experimental subject to name aggressive words (like "punch") under two conditions. In one condition the words were preceded by a non-weapon word like "rabbit" or "bug." In the second condition, the same words were preceded by a weapon word such as "gun" or "knife." The issue addressed

by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word, compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.



43.2, 42.9, 35.6, 35.1, 25.6, 25.4, 23.6, 20.5,
19.9, 14.4, 12.7, 11.3, 10.2, 10, 9.1, 7.5, 5.4,
4.7, 3.8, 2.1, 1.2, -0.2, -6.3, -6.7, -8.8, -10.4,
-10.5, -14.9, -14.9, -15, -18.5, -27.4

The effects of priming
(thousandths of a second).

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in the [link]. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf) the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number -27. The second-to-last row represents the numbers -10, -10, -15, etc. Once again, we have rounded the original values from [link].

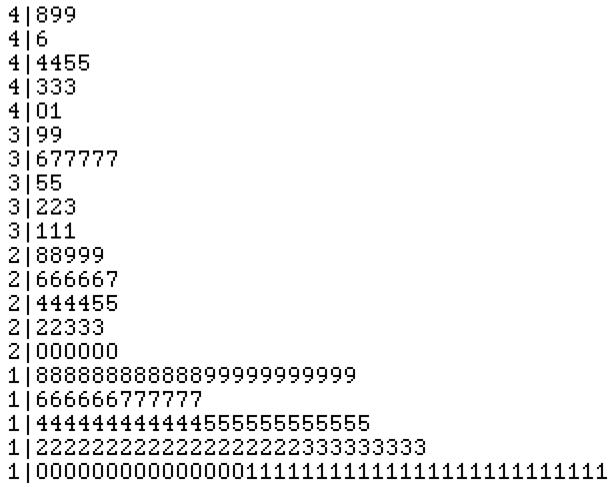
4|33
3|56

2|00456
1|00134
0|1245589
-0|0679
-1|005559
-2|7

Stem and leaf display with negative numbers and rounding

Observe that the figure contains a row headed by "0" and another headed by"-0". The stem of 0 is for numbers between 0 and 9 whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, -6, -7, and -9. Values that are exactly 0 before rounding should be split as evenly as possible between the "0" and "-0" rows. In [link], none of the values are 0 before rounding. The "0" that appears in the "-0" row comes from the original value of -0.2 in the figure.

Although stem and leaf displays are unwieldy for large datasets, they are often useful for datasets with up to 200 observations. [link] portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.



```
4|899
4|6
4|4455
4|333
4|01
3|99
3|677777
3|55
3|223
3|111
2|88999
2|666667
2|444455
2|22333
2|000000
1|88888888888899999999999
1|666666777777
1|444444444444555555555555
1|2222222222222222222333333333
1|00000000000000011111111111111111111111111
```

Stem and leaf display of populations of US cities with populations between 100,000 and 500,000.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example the largest number (493,559) was rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: 0-1, 2-3, 4-5, 6-7, and 8-9.

Whether your data can be suitably represented by a stem and leaf graph depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in [link] fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

Box Plots

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The **median**, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger. For example, consider the following data:

1 11.5 6 7.2 4 8 9 10 6.8 8.3 2 2 10 1

Ordered from smallest to largest:

1 1 2 2 4 6 **6.8** **7.2** 8 8.3 9 10 10 11.5

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.
Equation:

$$\frac{6.8+7.2}{2}=7$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of the data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

1 1 2 **2** 4 6 6.8

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is 9.

7.2 8 8.3 **9** 10 10 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

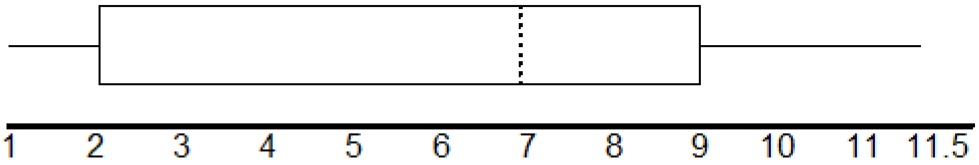
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

Note: You may encounter box and whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider the following data:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the TI web site):



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

Example:
The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77
Construct a box plot:
Using the TI-83, 83+, 84, 84+ Calculator

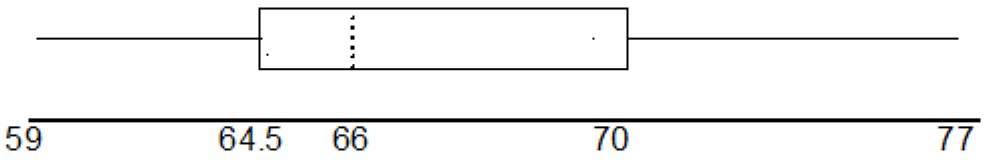
- Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, arrow down.
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
- Press ENTER
- Use the down and up arrow keys to scroll.

- Smallest value = 59
- Largest value = 77
- Q1: First quartile = 64.5

- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

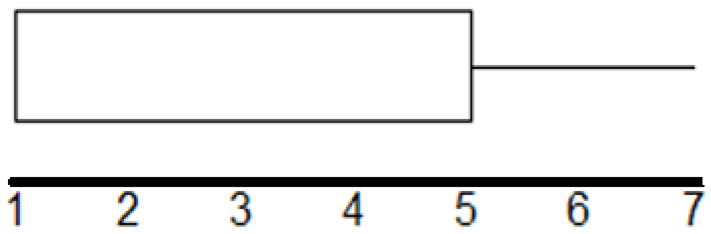
Using the TI-83, 83+, 84, 84+ to Construct the Box Plot
Go to 14:Appendix for Notes for the TI-83, 83+, 84, 84+ Calculator. To create the box plot:

- Press Y=. If there are any equations, press CLEAR to clear them.
- Press 2nd Y=.
- Press 4:Plotsoff. Press ENTER
- Press 2nd Y=
- Press 1:Plot1. Press ENTER.
- Arrow down and then use the right arrow key to go to the 5th picture which is the box plot. Press ENTER.
- Arrow down to Xlist: Press 2nd 1 for L1
- Arrow down to Freq: Press ALPHA. Press 1.
- Press ZOOM. Press 9:ZoomStat.
- Press TRACE and use the arrow keys to examine the box plot.



- **a**Each quarter has 25% of the data.
- **b**The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- **c**Interquartile Range: $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} = 70 - 64.5 = 5.5.$
- **d**The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- **e**The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example:
Test scores for a college statistics class held during the day are:
99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90
Test scores for a college statistics class held during the evening are:
98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85 79 78 98 90 79 81 25.5
Exercise:

Problem:

- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

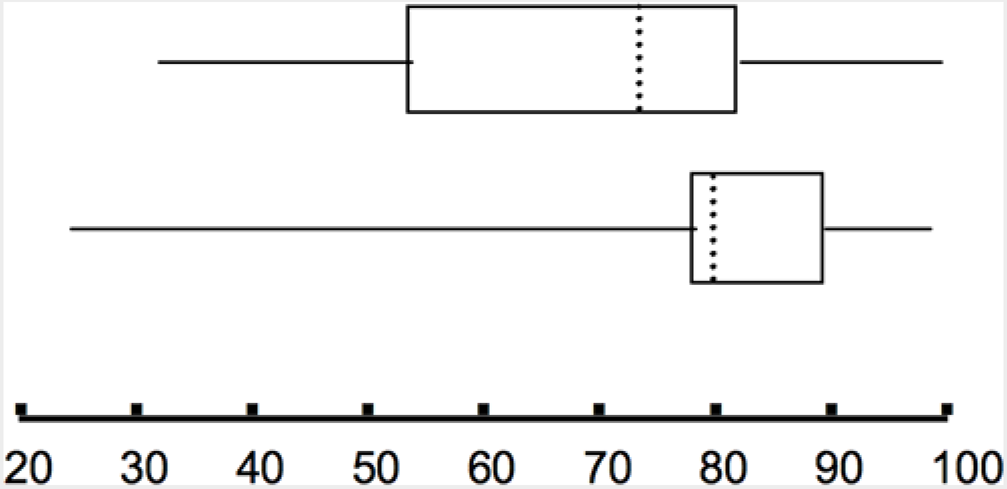
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

Solution:
First Data Set

- $\text{Xmin} = 32$
- $\text{Q1} = 56$
- $M = 74.5$
- $\text{Q3} = 82.5$
- $\text{Xmax} = 99$

Second Data Set

- $\text{Xmin} = 25.5$
- $\text{Q1} = 78$
- $M = 81$
- $\text{Q3} = 89$
- $\text{Xmax} = 98$



The first data set (the top box plot) has the widest spread for the middle 50% of the data. $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$ is $82.5 - 56 = 26.5$ for the first data set and $89 - 78 = 11$ for the second data set. So, the first set of data has its middle 50% of scores more spread out.
25% of the data is between M and $\mathrm{Q3}$ and 25% is between $\mathrm{Q3}$ and Xmax.

Glossary

Median
A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Quartiles
The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of the Location of the Data
Descriptive Statistics: Measuring the Location of Data explains percentiles and quartiles and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom contributed the section "Interpreting Percentiles, Quartile and the Median."

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3, is the same as the 75th percentile (75th %ile). The median, M, is called both the second quartile and the 50th percentile (50th %ile).

Note: Quartiles are given special attention in the Box Plots module in this chapter.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).
Equation:

$$\text{IQR} = Q_3 - Q_1$$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than** $(1.5)(\text{IQR})$ **below the first quartile or more than** $(1.5)(\text{IQR})$ **above the third quartile**. Potential outliers always need further investigation.

Example:
Exercise:

Problem:

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

389,950 230,500 158,000 479,000 639,000 114,950 5,500,000 387,000 659,000 529,000 575,000 488,800 1,095,000

Solution:

Order the data from smallest to largest.

114,950 158,000 230,500 387,000 389,950 479,000 488,800 529,000 575,000 639,000 659,000 1,095,000 5,500,000

$M = 488{,}800$

$Q_1 = \frac{230500+387000}{2} = 308750$

$Q_3 = \frac{639000+659000}{2} = 649000$

$\text{IQR} = 649000 - 308750 = 340250$

$(1.5)(\text{IQR}) = (1.5)(340250) = 510375$

$Q_1 - (1.5)(\text{IQR}) = 308750 - 510375 = -201625$

$Q_3 + (1.5)(\text{IQR}) = 649000 + 510375 = 1159375$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example:
Exercise:

Problem:

For the two data sets in the test scores example, find the following:

- **a**The interquartile range. Compare the two interquartile ranges.
- **b**Any outliers in either set.
- **c**The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Solution:

For the IQRs, see the answer to the test scores example. The first data set has the larger IQR, so the scores between Q3 and Q1 (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \ \cdot \ (\text{IQR}) \ = \ \left(\frac{3}{2}\right) \ \cdot \ (26.5) \ = \ 39.75$
- $\text{Xmax} \ \text{-} \ \text{Q3} \ = \ 99 \ \text{-} \ 82.5 \ = \ 16.5$
- $\text{Q1} \ \text{-} \ \text{Xmin} \ = \ 56 \ \text{-} \ 32 \ = \ 24$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $\text{Xmax} - \text{Q3} = 98 - 89 = 9$
- $\text{Q1} - \text{Xmin} = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and

Data Chapter). Get the percentiles from that chart.

First Data Set

- $30\text{th } \%\text{ile (between the 6th and 7th values)} = \frac{(56+59)}{2} = 57.5$
- $80\text{th } \%\text{ile (between the 16th and 17th values)} = \frac{(84+84.5)}{2} = 84.25$

Second Data Set

- $30\text{th } \%\text{ile (7th value)} = 78$
- $80\text{th } \%\text{ile (18th value)} = 90$

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Example:

Finding Quartiles and Percentiles Using a Table

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

Find the median: Look again at the "cumulative relative frequency " column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th %ile is between the 25th (7) and 26th (7) values. **The median is 7.**

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example:

Exercise:

Problem: Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile.
4. What is another name for the first quartile?

Solution:

1. $\frac{(8+9)}{2} \ = \ 8.5$
   Look where cum. rel. freq. = 0.80. 80% of the data is 8 or less. 80th %ile is between the last 8 and first 9.
2. 9
3. 6
4. First Quartile = 25th %ile

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct a table of the data.
4. Construct 2 different histograms. For each, starting value = _____ ending value = ____.
5. Use the table to find the median, first quartile, and third quartile.
6. Construct a box plot.
7. Use the table to find the following:
   - The 10th percentile
   - The 70th percentile
   - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. p% of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good'; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example:
On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example:
On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example:
At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles
Exercise:

Problem:

- **a** For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- **b** The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

- **c** A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Solution:

- **a** For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.
- **b**INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.
- **c**He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Exercise:

Problem:

- **a** For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- **b**The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- **a**For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
- **b**INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Solution:

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Exercise:

Problem:

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than you did. In this context, you would prefer a wait time corresponding to a lower percentile. INTERPRETATION: 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:

Problem:

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Solution:

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise:

Problem:

- The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
- b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Solution:

- **a**The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.
- **b** The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

**With contributions from Roberta Bloom

Glossary

Interquartile Range (IRQ)
The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

Outlier
An observation that does not fit the rest of the data.

Percentile
A number that divides ordered data into hundredths.

Example:
Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0...\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of Central Tendency

This chapter discusses measuring descriptive statistical information using the center of the data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts (previously discussed under box plots in this chapter). The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note: The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced "x bar"): x.

The Greek letter μ (pronounced "mew") represents the population mean. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

11122344444

Equation:

$$x=\frac{1+1+1+2+2+3+4+4+4+4+4}{11}=2.7$$

Equation:

$$x=\frac{3\times1+2\times2{+}1\times3+5\times4}{11}=2.7$$

In the second calculation for the sample mean, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2}$= $\frac{97+1}{2}$= 49. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2}$= $\frac{100+1}{2}$= 50.5. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

348810111213141515161617171821222224242526262727292931323333
33434353740444447

Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, **M**, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

34881011121314151516161717182122222424
2526262727292931323333343435374044 4447

$$M = \frac{24+24}{2} = 24$$

The median is 24.

Using the TI-83,83+,84, 84+ Calculators
Calculator Instructions are located in the menu item 14:Appendix (Notes for the TI-83, 83+, 84, 84+ Calculators).

- Enter data into the list editor. Press STAT 1:EDIT
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and ENTER.
- Press the down and up arrow keys to scroll.

$x = 23.6, M = 24$

Example:
Exercise:

Problem:

Suppose that, in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center," the mean or the median?

Solution:

$x=\frac{5000000+49\times 30000}{50}=129400$

$M=30000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example:
Statistics exam scores for 20 students are as follows
Statistics exam scores for 20 students are as follows:
50 53 59 59 63 63 72 72 72 72 72 76 78 81 83 84 84 84 90 93
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Example:
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note: The mode can be calculated for qualitative data as well as for quantitative data.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean x of the sample is very likely to get closer and closer to μ. This is discussed in more detail in **The Central Limit Theorem**.

Note: The formula for the mean is located in the Summary of Formulas section course.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean x is an example of a statistic which estimates the population mean μ.

Glossary

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Mode

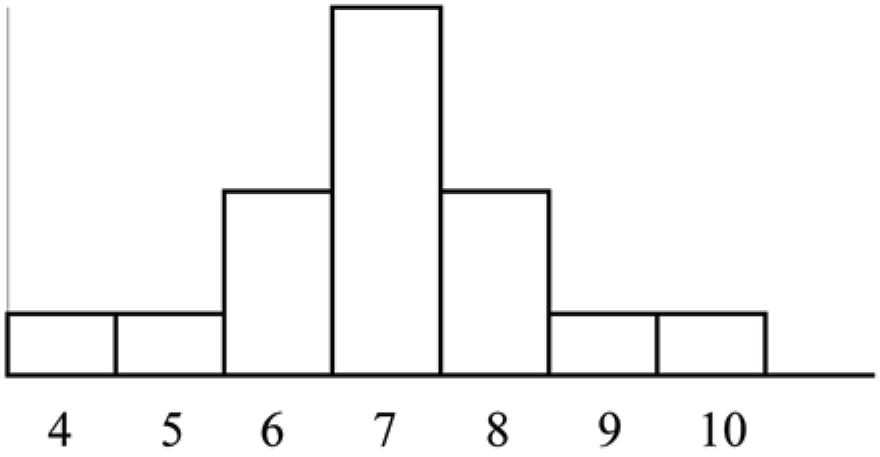
The value that appears most frequently in a set of data.

Skewness and the Mean, Median, and Mode

Consider the following data set:

4 5 6 6 6 7 7 7 7 7 7 8 8 8 9 10

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.

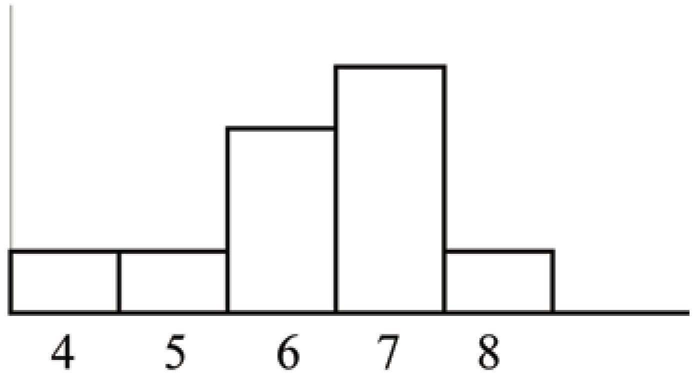


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 5 6 6 6 7 7 7 7 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

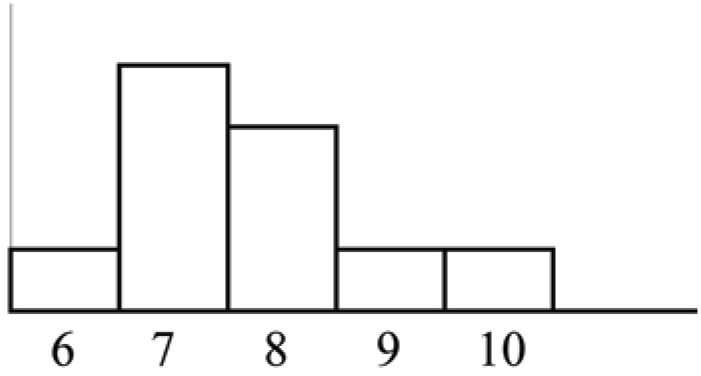


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 7 7 7 7 8 8 8 9 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Measures of the Spread of the Data
Descriptive Statistics: Measuring the Spread of Data explains standard deviation as a measure of variation in data and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom made contributions that helped to clarify the standard deviation and the variance.

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- can be used to determine whether a particular data value is close to or far from the mean

The standard deviation provides a measure of the overall variation in a data set
The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

Because market B has a higher standard deviation, we know that there is more variation in the waiting times at market B. Overall, wait times at market B are more spread out from the average; wait times at market A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.
Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

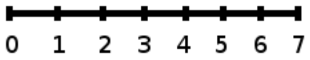
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is **one** standard deviation to the **right** of 5 because
$5 + (1)(2) = 7$.

If 1 were also part of the data set, then 1 is **two** standard deviations to the **left** of 5 because
$5 + (\text{-}2)(2) = 1$.



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7=5+**(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1=5+**(−2)**(2)

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population:

- **sample:** $x = x + (\#\text{ofSTDEV})(\text{s})$
- **Population:** $x = \mu + (\#\text{ofSTDEV})(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation
If x is a number, then the difference "x - mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - x$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is an **average of the squares of the deviations** (the $x - x$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n-1**, one less than the number of items in the sample. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f\cdot(x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f\cdot(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is 1. If a value appears three times in the data set or population, f is 3.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note: In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example:

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9 9.5 9.5 10 10 10 10 10.5 10.5 10.5 10.5 11 11 11 11 11 11 11.5 11.5 11.5

Equation:

$$x = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525$$

The average age is 10.53 years, rounded to 2 places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	**Freq.**	**Deviations**	Deviations^2	**(Freq.)**(Deviations^2**)**
x	f	$(x - x)$	$(x - x)^2$	$(f)(x - x)^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$

Data	**Freq.**	**Deviations**	Deviations^2	**(Freq.)**(Deviations^2**)**
11.5	*3*	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1):
$s^2 = \frac{9.7375}{20-1} = 0.5125$
The **sample standard deviation** s is equal to the square root of the sample variance:
$s = \sqrt{0.5125} = .0715891$ Rounded to two decimal places, $s = 0.72$
Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.
Exercise:

Problem: Verify the mean and standard deviation calculated above on your calculator or computer.

Solution:
Using the TI-83,83+,84+ Calculators

- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- x=10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: x = x + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = x + (#ofSTDEVs)(s) because the data is from a sample

Exercise:

Problem: Find the value that is 1 standard deviation above the mean. Find $(x + 1s)$.

Solution:

$(x + 1s) = 10.53 + (1)(0.72) = 11.25$

Exercise:

Problem: Find the value that is two standard deviations below the mean. Find $(x - 2s)$.

Solution:

$(x - 2s) = 10.53 - (2)(0.72) = 9.09$

Exercise:

Problem: Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

- $(x - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Explanation of the standard deviation calculation shown in the table
The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero**. (For this example, there are n=20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n=20, the calculation divided by n-1=20-1=19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n-1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n-1) gives a better estimate of the population variance.

Note: Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**.

Note: The formula for the standard deviation is at the end of the chapter.

Example:
Exercise:

Problem: Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

3342494953555561 6367686869697273 7478808388888890 929494949496100

- **a**Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- **b**Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
  - **i**The sample mean
  - **ii**The sample standard deviation
  - **iii**The median
  - **iv**The first quartile
  - **v**The third quartile
  - **vi**IQR
- **c**Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution:

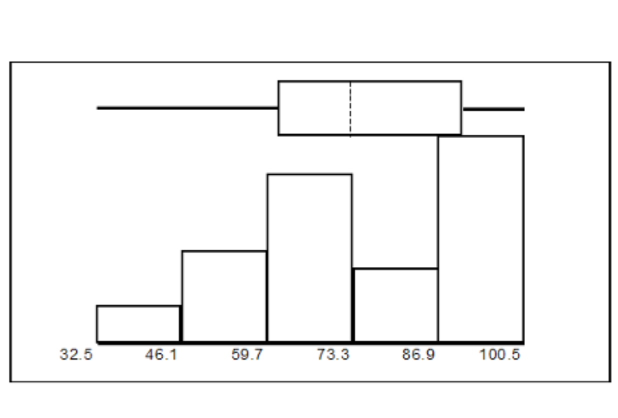
- **a**

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	**0.998** (Why isn't this value 1?)

- **b**
  - **i**The sample mean = 73.5
  - **ii**The sample standard deviation = 17.9
  - **iii**The median = 73
  - **iv**The first quartile = 61
  - **v**The third quartile = 90
  - **vi**IQR = 90 - 61 = 29

- **c**The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 - 33 = 40) than the spread in the upper 50% (100 - 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, it can be misleading to compare the data values directly.

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\text{ofSTDEVs} = \frac{\text{value}-\text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	x = x + z s	$z = \frac{x-x}{s}$
Population	x = μ + z σ	$z = \frac{x-\mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$\#\text{ofSTDEVs} = \frac{\text{value}-\text{mean}}{\text{standard deviation}}$; $z = \frac{x-\mu}{\sigma}$

For John, $z = \#\text{ofSTDEVs} = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ofSTDEVs} = \frac{77-80}{10} = -0.3$

John has the better G.P.A. when compared to his school because his G.P.A. is 0.21 standard deviations **below** his school's mean while Ali's G.P.A. is 0.3 standard deviations **below** his school's mean.

John's z-score of −0.21 is higher than Ali's z-score of −0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

Glossary

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variance

Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Summary of Formulas

A summary of useful formulas used in examining descriptive statistics

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- x = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x \quad f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x \quad f$ = The sum of values multiplied by their respective frequencies
- $x \qquad x$ or $x \qquad \mu$ = Deviations from the mean (how far a value is from the mean)
- $x \qquad x$ or $x \qquad \mu$ = Deviations squared
- $f \ x \qquad x$ or $f \ x \qquad \mu$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $x \qquad \frac{\sum x}{n}$ or $x \qquad \frac{\sum f\ x}{n}$
- $\mu \qquad \frac{\sum x}{N}$ or $\mu = \frac{\sum f\ x}{N}$

Standard Deviation Formulas:

- $s \qquad \overline{\frac{\Sigma\ x \quad x}{n}}$ or $s \qquad \overline{\frac{\Sigma f \quad x \quad x}{n}}$
- $\sigma \qquad \overline{\frac{\Sigma\ x \quad \mu}{N}}$ or $\sigma \qquad \overline{\frac{\Sigma f \quad x \quad \mu}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- x = x+ (#ofSTDEVs)(s)
- x = μ + (#ofSTDEVs)(σ)

Practice 1: Center of the Data
This module provides students with opportunities to apply concepts related to descriptive statistics. Students are asked to take a set of sample data and calculate a series of statistical values for that data.

Student Learning Outcomes

- The student will calculate and interpret the center, spread, and location of the data.
- The student will construct and interpret histograms an box plots.

Given

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Complete the Table

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Discussion Questions

Exercise:

Problem: What does the frequency column sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column sum to? Why?

Solution:

1

Exercise:

Problem:

What is the difference between relative frequency and frequency for each data value?

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

Enter the Data

Enter your data into your calculator or computer.

Construct a Histogram

Determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram below. Label the horizontal and vertical axes with words. Include numerical scaling.

Data Statistics

Calculate the following values:

Exercise:

Problem: Sample mean = $\overline{x}$ =

Solution:

4.75

Exercise:

Problem: Sample standard deviation = s_x =

Solution:

1.39

Exercise:

Problem: Sample size = n =

Solution:

65

Calculations

Use the table in section 2.11.3 to calculate the following values:
Exercise:

Problem: Median =

Solution:

4

Exercise:

Problem: Mode =

Solution:

4

Exercise:

Problem: First quartile =

Solution:

4

Exercise:

Problem: Second quartile = median = 50th percentile =

Solution:

4

Exercise:

Problem: Third quartile =

Solution:

6

Exercise:

Problem: Interquartile range (IQR) = _____ - _____ = _____

Solution:

$6 - 4 = 2$

Exercise:

Problem: 10th percentile =

Solution:

3

Exercise:

Problem: 70th percentile =

Solution:

6

Exercise:

Problem: Find the value that is 3 standard deviations:

- **a**Above the mean
- **b**Below the mean

Solution:

- **a**8.93

- **b**0.58

Box Plot

Construct a box plot below. Use a ruler to measure and scale accurately.

Interpretation

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Practice 2: Spread of the Data
Practice exercise for Descriptive Statistics

Student Learning Outcomes

- The student will calculate measures of the center of the data.
- The student will calculate the spread of the data.

Given

The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976-77 through 2004-2005. (*Source: Graphically Speaking by Bill King, LTCC Institutional Research, December 2005*).

Use these values to answer the following questions:

- μ = 1000 FTES
- Median = 1014 FTES
- σ = 474 FTES
- First quartile = 528.5 FTES
- Third quartile = 1447.5 FTES
- n = 29 years

Calculate the Values

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

6

Exercise:

Problem: 75% of all years have a FTES:

- **a**At or below:
- **b**At or above:

Solution:

- **a**1447.5
- **b**528.5

Exercise:

Problem: The population standard deviation =

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Solution:

50%

Exercise:

Problem: What is the IQR? What does the IQR represent?

Solution:

919

Exercise:

Problem:

How many standard deviations away from the mean is the median?

Solution:

0.03

Additional Information: The population FTES for 2005-2006 through 2010-2011 was given in an updated report. (Source: http://www.ltcc.edu/data/ResourcePDF/LTCC_FactBook_2010-11.pdf). The data are reported here.

Year	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11
Total FTES	1585	1690	1735	1935	2021	1890

Exercise:

Problem:

Calculate the mean, median, standard deviation, first quartile, the third quartile and the IQR. Round to one decimal place.

Solution:

mean = 1809.3
median = 1812.5
standard deviation = 151.2
First quartile = 1690

Third quartile = 1935
IQR = 245

Exercise:

Problem:

Construct a boxplot for the FTES for 2005-2006 through 2010-2011 and a boxplot for the FTES for 1976-1977 through 2004-2005.

Exercise:

Problem:

Compare the IQR for the FTES for 1976-77 through 2004-2005 with the IQR for the FTES for 2005-2006 through 2010-2011. Why do you suppose the IQRs are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Homework

Descriptive Statistics: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean and provides homework questions related to lessons about descriptive statistics.

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

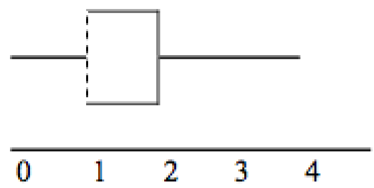
# of movies	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0	5		
1	9		
2	6		
3	4		
4	1		

- **a**Find the sample mean $\overline{x}$
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.
- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.

- **i**What percent of the students saw fewer than three movies?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data.
- **m**Construct a stem plot of the data.

Solution:

- **a**1.48
- **b**1.12
- **e**1
- **f**1
- **g**2
- **h**



- **i**80%
- **j**1
- **k**3

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years. ((*Source: http://www.usatoday.com/news/nation/story/2012-05-17/minority-births-census/55029100/1)*)

- **a**Based upon this information, give two reasons why the black median age could be lower than the white median age.
- **b**Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

- **c**How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

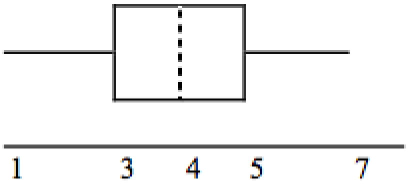
X	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	5		
3	8		
4	12		
5	12		
7	1		

- **a**Find the sample mean $\overline{x}$
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.

- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.
- **i**What percent of the students owned at least five pairs?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data
- **m**Construct a stem plot of the data

Solution:

- **a**3.78
- **b**1.29
- **e**3
- **f**4
- **g**5
- **h**



- **i**32.5%
- **j**4
- **k**5

Exercise:

Problem:

600 adult Americans were asked by telephone poll, What do you think constitutes a middle-class income? The results are below. Also, include left endpoint, but not the right endpoint. (*Source: Time magazine; survey by Yankelovich Partners, Inc.*)

Note:"Not sure" answers were omitted from the results.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000 - 25,000	0.09
25,000 - 30,000	0.19
30,000 - 40,000	0.26
40,000 - 50,000	0.18
50,000 - 75,000	0.17
75,000 - 99,999	0.02
100,000+	0.01

- **a**What percent of the survey answered "not sure" ?
- **b**What percent think that middle-class is from $25,000 - $50,000 ?
- **c**Construct a histogram of the data

    1. **i**Should all bars have the same width, based on the data? Why or why not?
    2. **ii**How should the <20,000 and the 100,000+ intervals be handled? Why?

- **d**Find the 40th and 80th percentiles
- **e**Construct a bar graph of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year (*Source: San Jose Mercury News*)

177 205 210 210 232 205 185 185 178 210 206 212 184 174 185 242 188 212 215 247 241 223 220 260 245 259 278 270 280 295 275 285 290 272 273 280 285 286 200 215 185 230 250 241 190 260 250 302 265 290 276 228 265

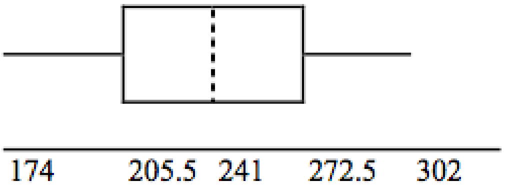
- **a**Organize the data from smallest to largest value.
- **b**Find the median.
- **c**Find the first quartile.
- **d**Find the third quartile.
- **e**Construct a box plot of the data.
- **f**The middle 50% of the weights are from _______ to _______.
- **g**If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- **h**If our population were the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- **i**Assume the population was the San Francisco 49ers. Find:

  - **i**the population mean, μ.
  - **ii**the population standard deviation, σ.
  - **iii**the weight that is 2 standard deviations below the mean.
  - **iv**When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

- **j**That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution:

- **b**241
- **c**205.5
- **d**272.5
- **e**



- **f**205.5, 272.5
- **g**sample
- **h**population
- **i**
    - **i**236.34
    - **ii**37.50
    - **iii**161.34
    - **iv**0.84 std. dev. below the mean
- **j**Young

Exercise:

Problem:

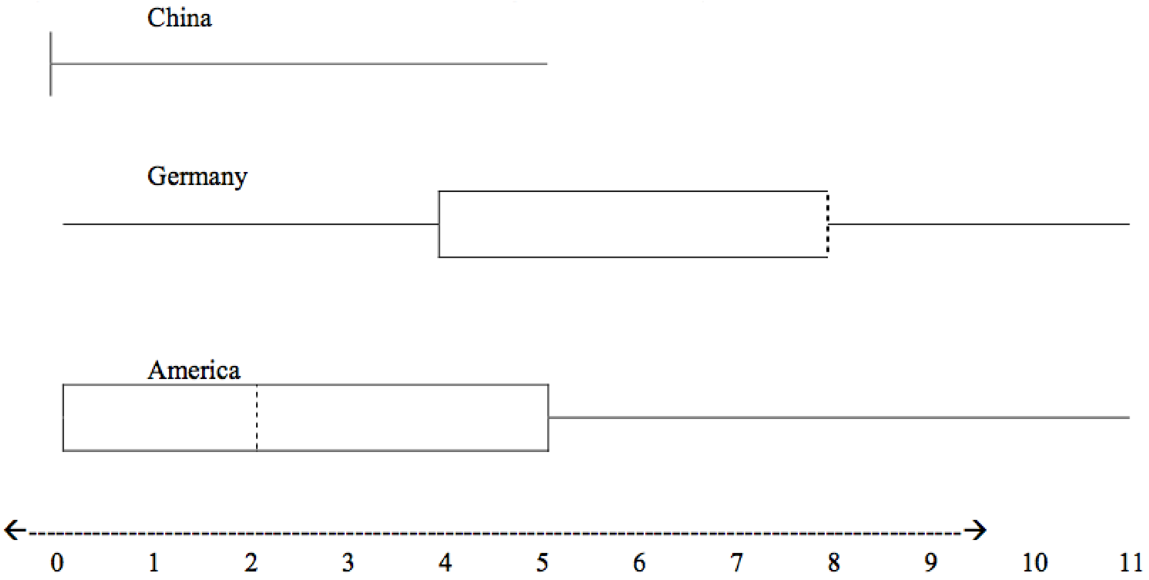
An elementary school class ran 1 mile with a mean of 11 minutes and a standard deviation of 3 minutes. Rachel, a student in the class, ran 1 mile in 8 minutes. A junior high school class ran 1 mile with a mean of 9 minutes and a standard deviation of 2 minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran 1 mile with a mean of 7 minutes and a standard deviation of 4 minutes. Nedda, a student in the class, ran 1 mile in 8 minutes.

- **a**Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
- **b**Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

In a survey of 20 year olds in China, Germany and America, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
- **b**Explain how it is possible that more Americans than Germans surveyed have been to over eight foreign countries.
- **c**Compare the three box plots. What do they imply about the foreign travel of twenty year old residents of the three countries when compared to each other?

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The twelve change scores are as follows:

3 8 -1 2 0 5 -3 1 -1 6 5 -2

- **a**What is the mean change score?
- **b**What is the standard deviation for this population?
- **c**What is the median change score?
- **d**Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best G.P.A. when compared to his school? Explain how you determined your answer.

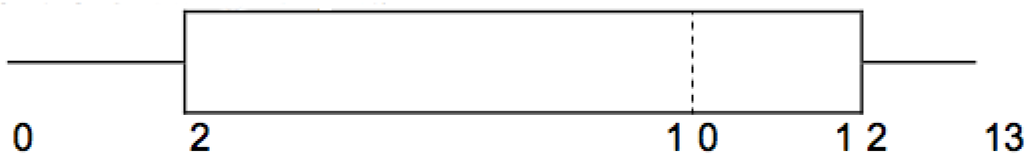
Student	G.P.A.	School Ave. G.P.A.	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Solution:

Kamala

Exercise:

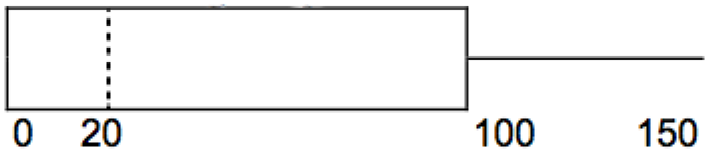
Problem: Given the following box plot:



- **a**Which quarter has the smallest spread of data? What is that spread?
- **b**Which quarter has the largest spread of data? What is that spread?
- **c**Find the Inter Quartile Range (IQR).
- **d**Are there more data in the interval 5 - 10 or in the interval 10 - 13? How do you know this?
- **e**Which interval has the fewest data in it? How do you know this?
  - **I** 0-2
  - **II**2-4
  - **III**10-12
  - **IV**12-13

Exercise:

Problem: Given the following box plot:



- **a**Think of an example (in words) where the data might fit into the above box plot. In 2-5 sentences, write down the example.
- **b**What does it mean to have the first and second quartiles so close together, while the second to fourth quartiles are far apart?

Exercise:

Problem:

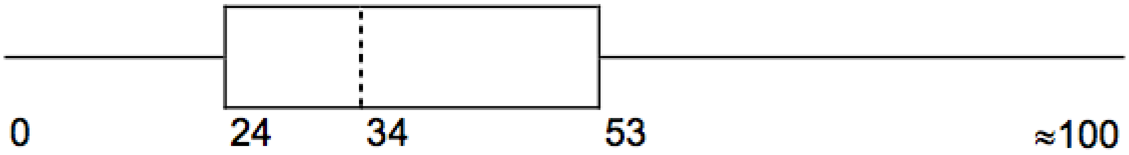
Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows. (*Source: West magazine*)

Age Group	**Percent of Community**
0-17	18.9
18-24	8.0
25-34	22.8
35-44	15.0
45-54	13.1

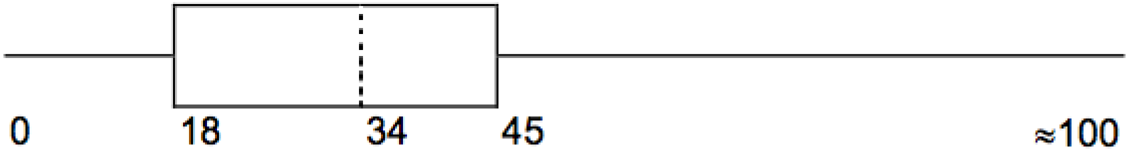
Age Group	**Percent of Community**
55-64	11.9
65+	10.3

- **a**Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not?
- **b**What percent of the community is under age 35?
- **c**Which box plot most resembles the information above?

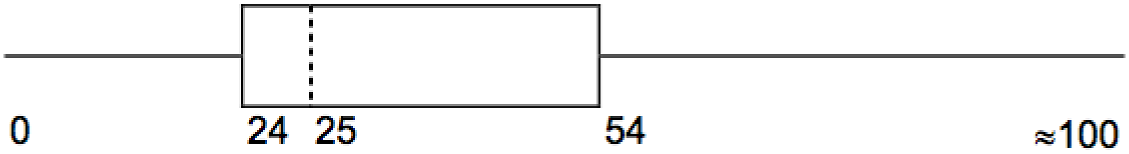
i.



ii.



iii.



Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, each asked adult consumers the number of fiction paperbacks they had purchased the previous month. The results are below.

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	

# of books	**Freq.**	**Rel. Freq.**
5	10	
7	5	
9	1	

Publisher B

# of books	**Freq.**	**Rel. Freq.**
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

Publisher C

- **a**Find the relative frequencies for each survey. Write them in the charts.
- **b**Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
- **c**In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.

- **d**Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
- **e**Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
- **f**Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all on-board transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Below is a summary of the bills for each group.

Amount($)	**Frequency**	**Rel. Frequency**
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

Singles

Amount($)	**Frequency**	**Rel. Frequency**
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

Couples

- **a**Fill in the relative frequency for each group.
- **b**Construct a histogram for the Singles group. Scale the x-axis by $50. widths. Use relative frequency on the y-axis.
- **c**Construct a histogram for the Couples group. Scale the x-axis by $50. Use relative frequency on the y-axis.
- **d**Compare the two graphs:

  - **i**List two similarities between the graphs.
  - **ii**List two differences between the graphs.
  - **iii**Overall, are the graphs more similar or different?

- **e**Construct a new graph for the Couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
- **f**Compare the graph for the Singles with the new graph for the Couples:

  - **i**List two similarities between the graphs.
  - **ii**Overall, are the graphs more similar or different?

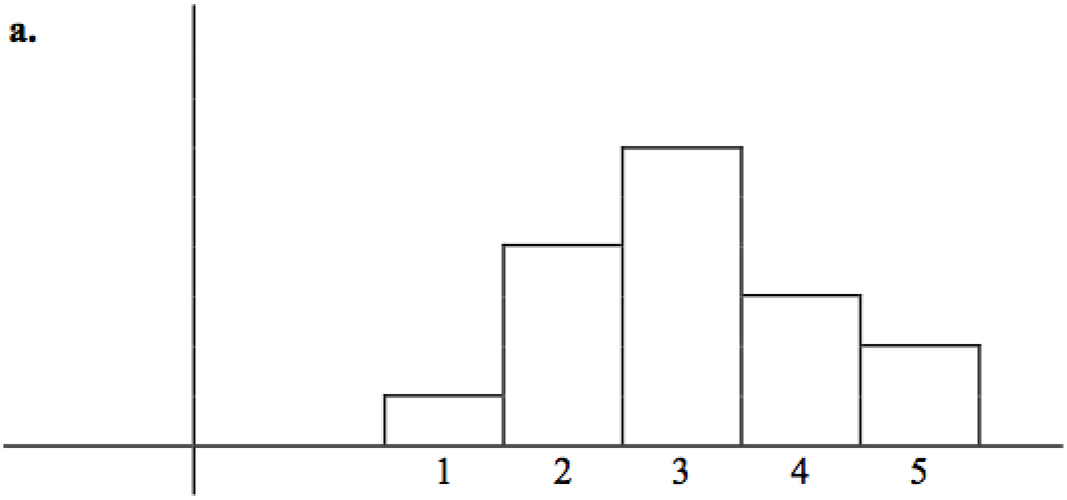
- **i**By scaling the Couples graph differently, how did it change the way you compared it to the Singles?
- **j**Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person in a couple? Explain why in one or two complete sentences.

Exercise:

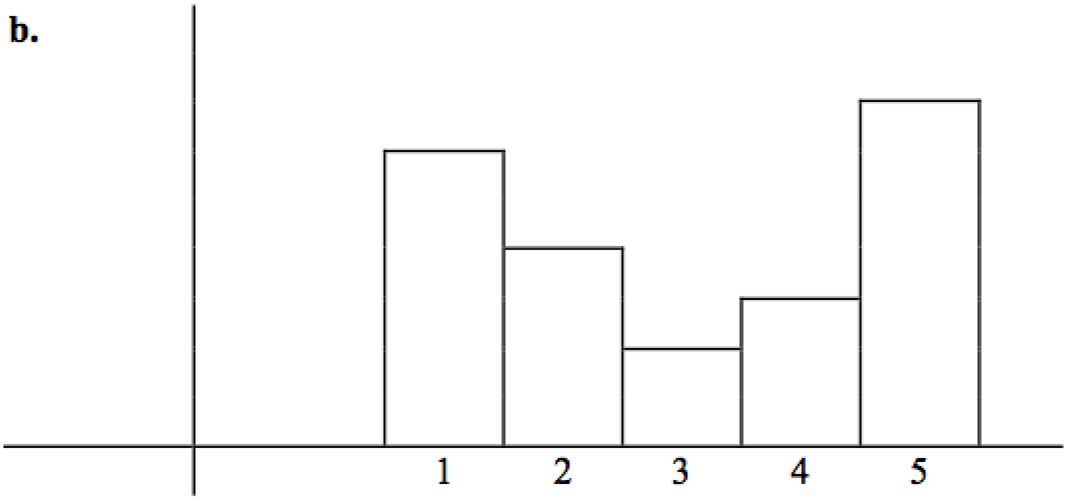
Problem:

Refer to the following histograms and box plot. Determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

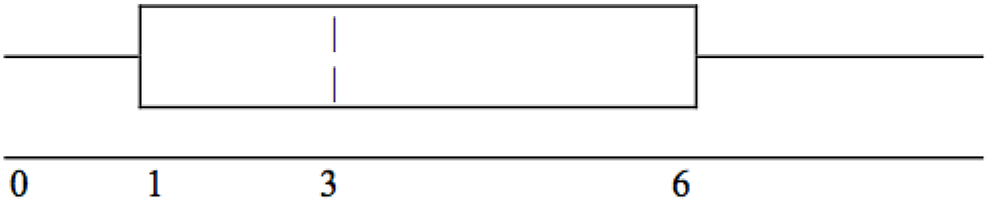
a.



b.



c.



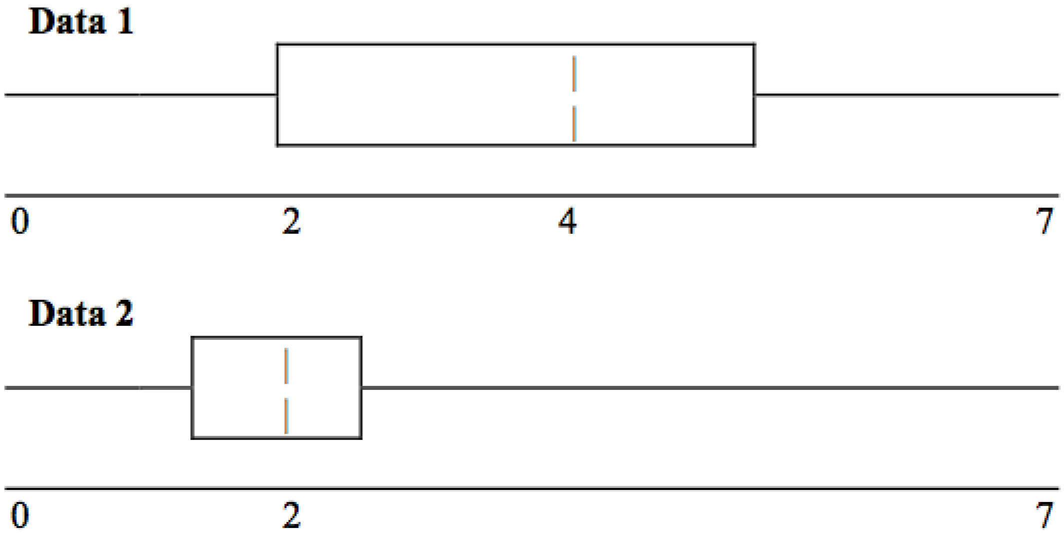
- **a**The medians for all three graphs are the same.
- **b**We cannot determine if any of the means for the three graphs is different.
- **c**The standard deviation for (b) is larger than the standard deviation for (a).
- **d**We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

- **a**True
- **b**True
- **c**True
- **d**False

Exercise:

Problem: Refer to the following box plots.



- **a**In complete sentences, explain why each statement is false.
    - **i****Data 1** has more data values above 2 than **Data 2** has above 2.
    - **ii**The data sets cannot have the same mode.
    - **iii**For **Data 1**, there are more data values below 4 than there are above 4.
- **b**For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences

Exercise:

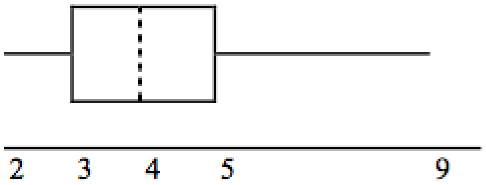
Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- **a**Organize the data in a chart.
- **b**Find the median, the first quartile, and the third quartile.
- **c**Find the 65th percentile.
- **d**Find the 10th percentile.
- **e**Construct a box plot of the data.
- **f**The middle 50% of the conferences last from _______ days to _______ days.
- **g**Calculate the sample mean of days of engineering conferences.
- **h**Calculate the sample standard deviation of days of engineering conferences.
- **i**Find the mode.
- **j**If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- **k**Give two reasons why you think that 3 - 5 days seem to be popular lengths of engineering conferences.

Solution:

- **b**4,3,5
- **c**4
- **d**3
- **e**



- **f**3,5
- **g**3.94
- **h**1.28
- **i**3
- **j**mode

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures (*source: Microsoft Bookshelf*):

6414 1550 2109 9350 21828 4300 5944 5722 2825 2044 5481 5200 5853 2750 10012 6357 27000 9414 7681 3200 17500 9200 7380 18314 6557 13713 17768 7493 2771 2861 1263 7285 28165 5080 11622

- **a**Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
- **b**Construct a histogram of the data.
- **c**If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- **d**Calculate the sample mean.
- **e**Calculate the sample standard deviation.
- **f**A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years. (*Source: Bureau of the Census*)

- **a**What does it mean for the median age to rise?
- **b**Give two reasons why the median age could rise.
- **c**For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

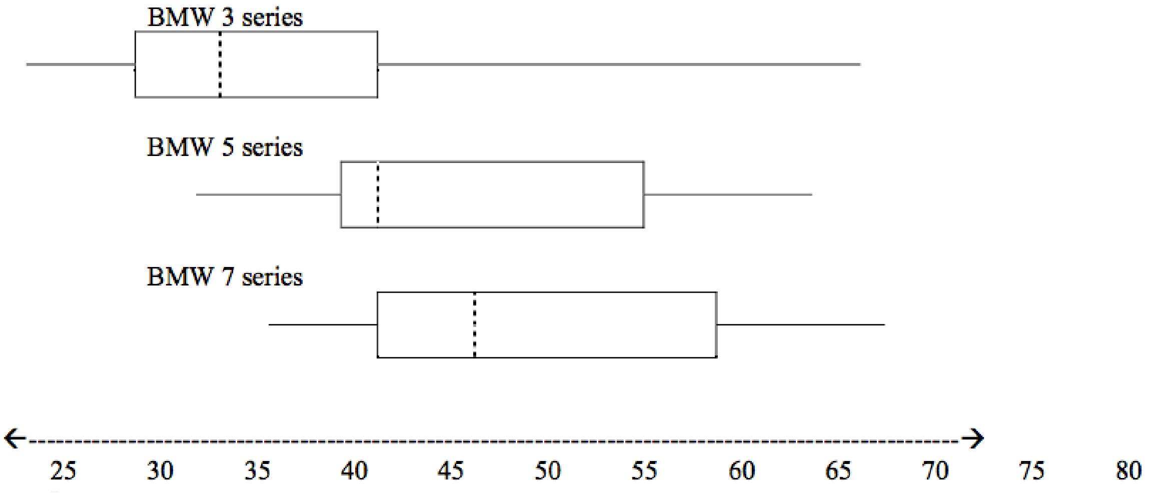
Solution:

- **c**Maybe

Exercise:

Problem:

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.

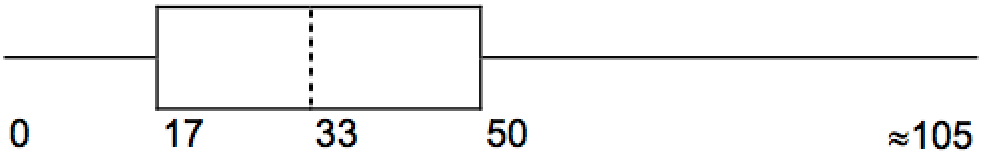
- **b**Which group is most likely to have an outlier? Explain how you determined that.
- **c**Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
- **d**Look at the BMW 5 series. Which quarter has the smallest spread of data? What is that spread?
- **e**Look at the BMW 5 series. Which quarter has the largest spread of data? What is that spread?
- **f**Look at the BMW 5 series. Estimate the Inter Quartile Range (IQR).
- **g**Look at the BMW 5 series. Are there more data in the interval 31-38 or in the interval 45-55? How do you know this?
- **h**Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

    - **i**31-35
    - **ii**38-41
    - **iii**41-64

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year. (Source: Bureau of the Census, 1990 Census)



- **a**Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
- **b**12.6% are age 65 and over. Approximately what percent of the population are of working age adults (above age 17 to age 65)?

Solution:

- **a**more children
- **b**62.4%

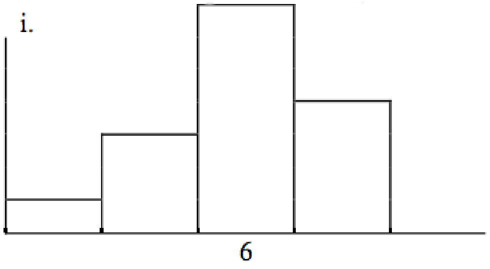
Exercise:

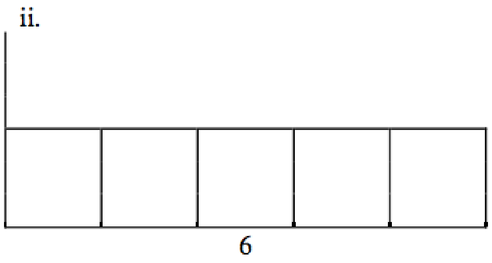
Problem:

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

	Javier	**Ercilla**
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

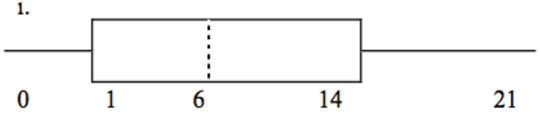
- **a**How can you determine which survey was correct ?
- **b**Explain what the difference in the results of the surveys implies about the data.
- **c**If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



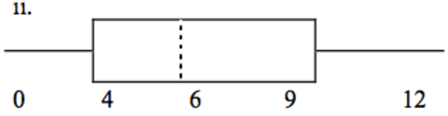


- **d**If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?

i.



ii.



Exercise:

Problem: Student grades on a chemistry exam were:

77, 78, 76, 81, 86, 51, 79, 82, 84, 99

- **a**Construct a stem-and-leaf plot of the data.
- **b**Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Solution:

- **b**51,99

Try these multiple choice questions (Exercises 24 - 30).

The next three questions refer to the following information. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency

Number of years	**Frequency**
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

Exercise:

Problem: What is the IQR?

- **A**8
- **B**11
- **C**15
- **D**35

Solution:

A

Exercise:

Problem: What is the mode?

- **A**19
- **B**19.5
- **C**14 and 20
- **D**22.65

Solution:

A

Exercise:

Problem: Is this a sample or the entire population?

- **A**sample
- **B**entire population
- **C**neither

Solution:

B

The next two questions refer to the following table. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise:

Problem: The 80th percentile is:

- **A**5
- **B**80
- **C**3
- **D**4

Solution:

D

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately:

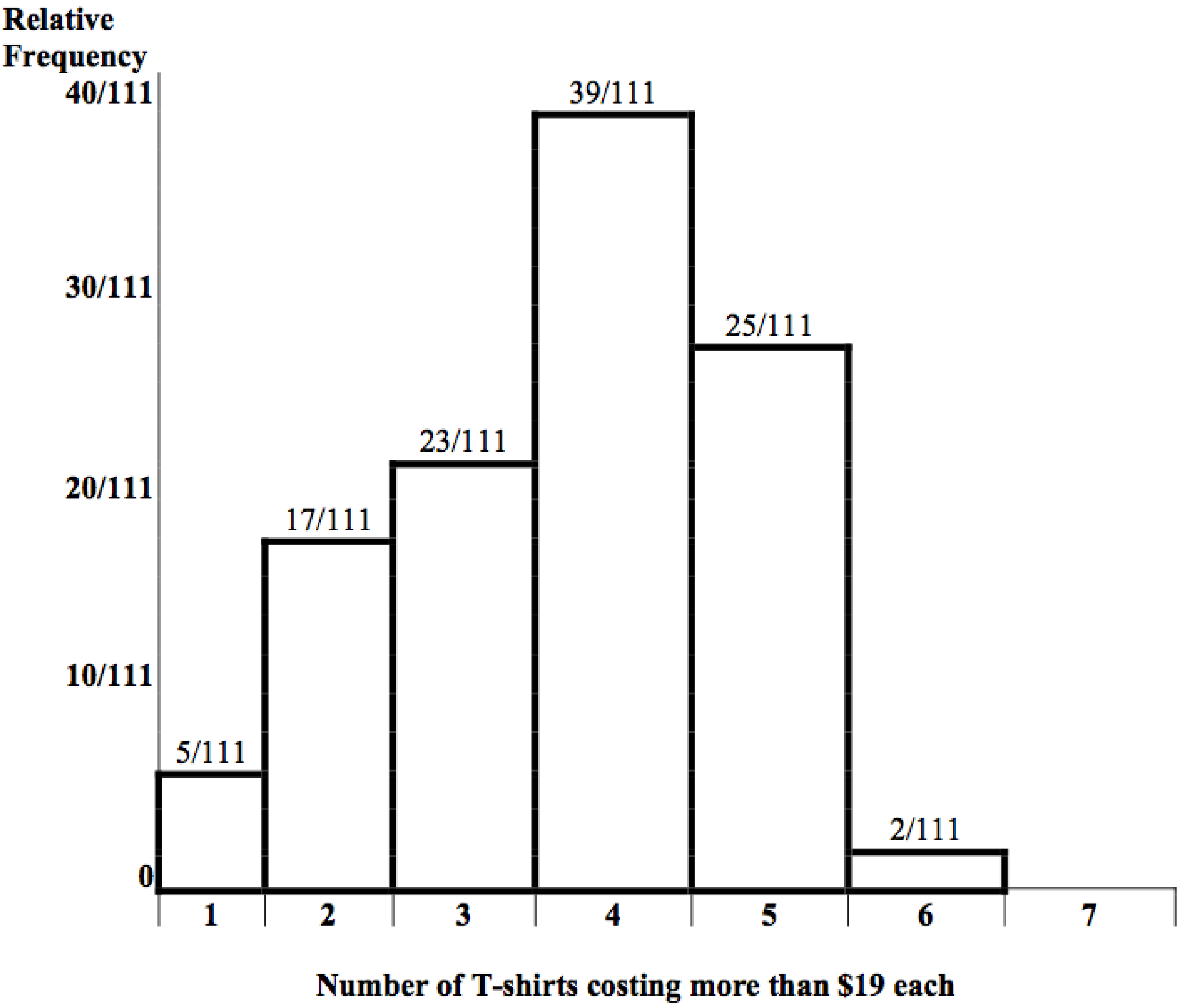
- **A**0.7

- **B**4.8
- **C**-2.8
- **D**Cannot be determined

Solution:

A

The next two questions refer to the following histogram. Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



Exercise:

Problem:

The percent of people that own at most three (3) T-shirts costing more than $19 each is approximately:

- **A**21
- **B**59
- **C**41
- **D**Cannot be determined

Solution:

C

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

- **A**cluster
- **B**simple random
- **C**stratified
- **D**convenience

Solution:

D

Exercise:

Problem:

Below are the **2010 obesity rates by U.S. states and Washington, DC.**(*Source: http://www.cdc.gov/obesity/data/adult.html)*)

State	Percent (%)	State	Percent (%)
Alabama	32.2	Montana	23.0
Alaska	24.5	Nebraska	26.9
Arizona	24.3	Nevada	22.4
Arkansas	30.1	New Hampshire	25.0
California	24.0	New Jersey	23.8
Colorado	21.0	New Mexico	25.1
Connecticut	22.5	New York	23.9
Delaware	28.0	North Carolina	27.8
Washington, DC	22.2	North Dakota	27.2
Florida	26.6	Ohio	29.2
Georgia	29.6	Oklahoma	30.4
Hawaii	22.7	Oregon	26.8
Idaho	26.5	Pennsylvania	28.6
Illinois	28.2	Rhode Island	25.5
Indiana	29.6	South Carolina	31.5

State	Percent (%)	State	Percent (%)
Iowa	28.4	South Dakota	27.3
Kansas	29.4	Tennessee	30.8
Kentucky	31.3	Texas	31.0
Louisiana	31.0	Utah	22.5
Maine	26.8	Vermont	23.2
Maryland	27.1	Virginia	26.0
Massachusetts	23.0	Washington	25.5
Michigan	30.9	West Virginia	32.5
Minnesota	24.8	Wisconsin	26.3
Mississippi	34.0	Wyoming	25.1
Missouri	30.5		

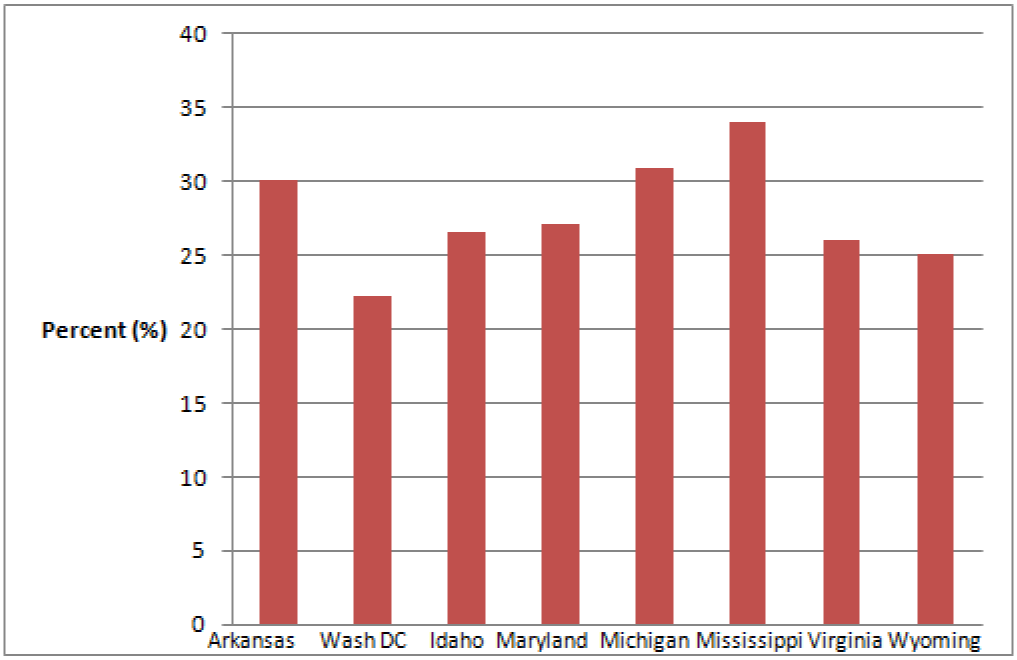
- **a.**Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.
- **b.**Use a random number generator to randomly pick 8 states. Construct a bar graph of the obesity rates of those 8 states.
- **c.**Construct a bar graph for all the states beginning with the letter "A."
- **d.**Construct a bar graph for all the states beginning with the letter "M."

Solution:

Example solution for **b** using the random number generator for the Ti-84 Plus to generate a simple random sample of 8 states. Instructions are below.

- Number the entries in the table 1 - 51 (Includes Washington, DC; Numbered vertically)
- Press MATH
- Arrow over to PRB
- Press 5:randInt(
- Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}. Corresponding percents are {28.7 21.8 24.5 26 28.9 32.8 25 24.6}.



Exercise:

Problem:

A music school has budgeted to purchase 3 musical instruments. They plan to purchase a piano costing $3000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer numerically.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar cost the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the table below. (Note that this is the data presented for publisher B in homework exercise 13).

# of books	Freq.	Rel. Freq.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

a. Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than 2 standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts (a) and (c) of this problem give the same answer?
e. Examine the shape of the data. Which part, (a) or (c), of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Solution:

- IQR = 4 – 1 = 3 ; Q1 – 1.5*IQR = 1 – 1.5(3) = -3.5 ; Q3 + 1.5*IQR = 4 + 1.5(3) = 8.5 ;The data value of 9 is larger than 8.5. The purchase of 9 books in one month is an outlier.
- The outlier should be investigated to see if there is an error or some other problem in the data; then a decision whether to include or exclude it should be made based on the particular situation. If it was a correct value then the data value should remain in the data set. If there is a problem with this data value, then it should be corrected or removed from the data. For example: If the data was recorded incorrectly (perhaps a 9 was miscoded and the correct value was 6) then the data should be corrected. If it was an error but the correct value is not known it should be removed from the data set.
- xbar – 2s = 2.45 – 2*1.88 = -1.31 ; xbar + 2s = 2.45 + 2*1.88 = 6.21 ; Using this method, the five data values of 7 books purchased and the one data value of 9 books purchased would be considered unusual.
- No: part (a) identifies only the value of 9 to be an outlier but part (c) identifies both 7 and 9.
- The data is skewed (to the right). It would be more appropriate to use the method involving the IQR in part (a), identifying only the one value of 9 books purchased as an outlier. Note that part (c) remarks that identifying unusual data values by using the criteria of being further than 2 standard deviations away from the mean is most appropriate when the data are mound-shaped and symmetric.
- The data are skewed to the right. For skewed data it is more appropriate to use the median as a measure of center.

**Exercises 32 and 33 contributed by Roberta Bloom

Introduction to Bivariate Data

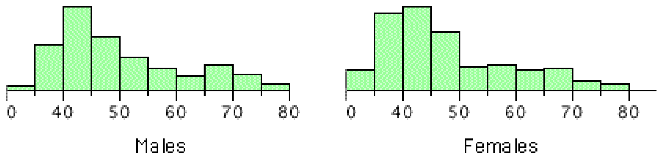
Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two **quantitative variables** for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. It helps to discuss something familiar since knowing the subject matter goes a long way in making judgments about statistical results. Let's begin by asking if people tend to marry other people of about the same age. Our experience tells us "yes," but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table 1 below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big surprise, but at least the data bear out our experiences, which is not always the case.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Sample of spousal ages of 10 White American Couples.

The pairs of ages in [link] are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a **histogram** (see [link]) and by a mean and standard deviation (See [link]).



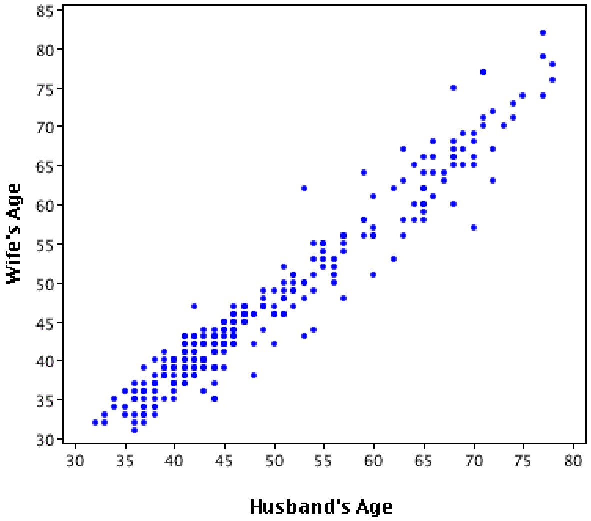
Histograms of spousal ages.

	Mean	Standard Deviation
Husband	49	11
Wife	47	11

Means and standard deviations of spousal ages.

Each distribution is fairly skewed with a long right tail. From [link] we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples have younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

We can learn much more by displaying the **bivariate** data in a graphical form that maintains the pairing. [link] shows a **scatter plot** of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

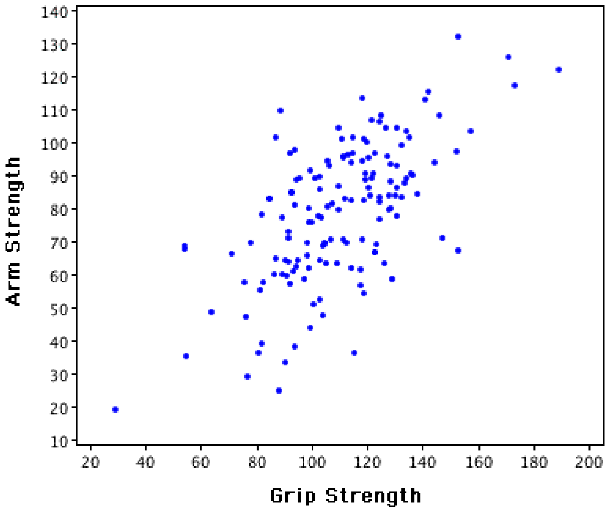


Scatterplot showing wife age as a function of husband age.

There are two important characteristics of the data revealed by [link]. First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (y) increases with the second variable (v), we say that x and y have a **positive association**. Conversely, when y decreases as x increases, we say that they have a **negative association**.

Second, the points cluster along a straight line. When this occurs, the relationship is called a **linear relationship**.

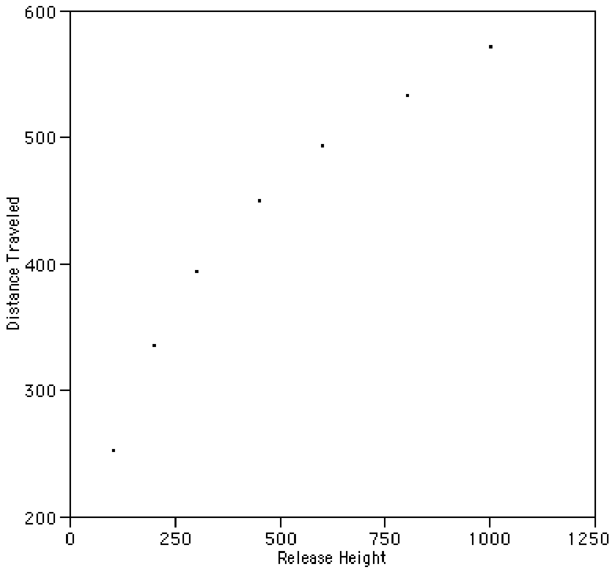
[link] shows a scatterplot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.



Scatter plot of Grip Strength and Arm Strength.

Not all scatter plots show linear relationships. [link] shows the results of an experiment conducted by Galileo on projectile motion. In the experiment, Galileo rolled balls down incline and measured how far they traveled as a function of the release height. It is clear from [link] that the relationship between "Release Height" and "Distance Traveled" is not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola.

- D. Dickey and T. Arnold's description of the study including a movie
- Rice University's Galilieo Project, section by S. Jennings



Galileo's data showing a non-linear relationship.

Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line. A statistical measure of the strength of the relationship between variables that takes these factors into account is the subject of the next section.

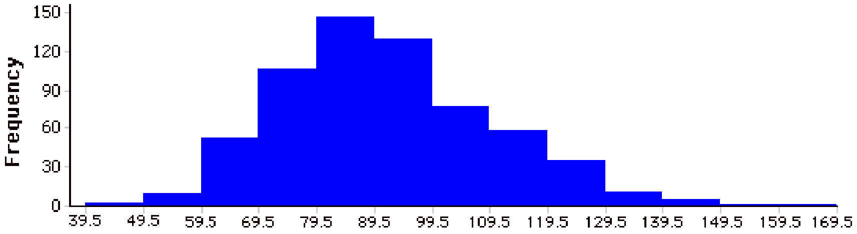
Glossary

Quantitative Variables

Variables that have are measured on a numeric or quantitative scale. Ordinal, interval and ratio scales are quantitative. A country's population, a person's shoe size, or a car's speed are all quantitative variables. Variables that are not quantitative are known as qualitative variables.

Histogram

A histogram is a graphical representation of a distribution. It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the bars represent the class frequencies.



See also: **Sturgis's Rule**

Sturgis's Rule

One method of determining the number of classes for a **histogram**, Sturgis's Rule is to take $1 + \log_2 N$ classes, rounded to the nearest integer.

Bivariate

Bivariate data is data for which there are two variables for each observation. As an example, the following bivariate data show the ages of husbands and wives of 10 married couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables. The scatter plot shown in [link] illustrates data from one of Galileo's classic experiments in which he observed the distance traveled balls traveled after being dropped on a incline as a function of their release height.

Positive Association

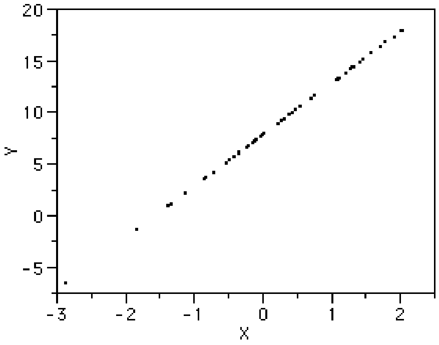
There is a positive association between variables X and Y if smaller values of X are associated with smaller values of Y and larger values of X are assoicated with larger values of Y.

Negative Association

There is a negative association between variables X and Y if smaller values of X are associated with larger values of Y and larger values of X are assoicated with smaller values of Y.

Linear Relationship

If the relationship between two variables is a perfect linear relationship, then a scatterplot of the points will fall on a straight line as shown in [link].



With real data, there is almost never a perfect linear relationship between two variables. The more the points tend to fall along a straight line the stronger the linear relationship. [link] shows two variables (husband's age and wife's age) that have a strong but not a perfect linear relationship.

Linear Regression and Correlation: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:

Equation:

$$y=a+\mathrm{bx}$$

where a and b are constant numbers.

x is the independent variable, and y is the dependent variable. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:

The following examples are linear equations.

Equation:

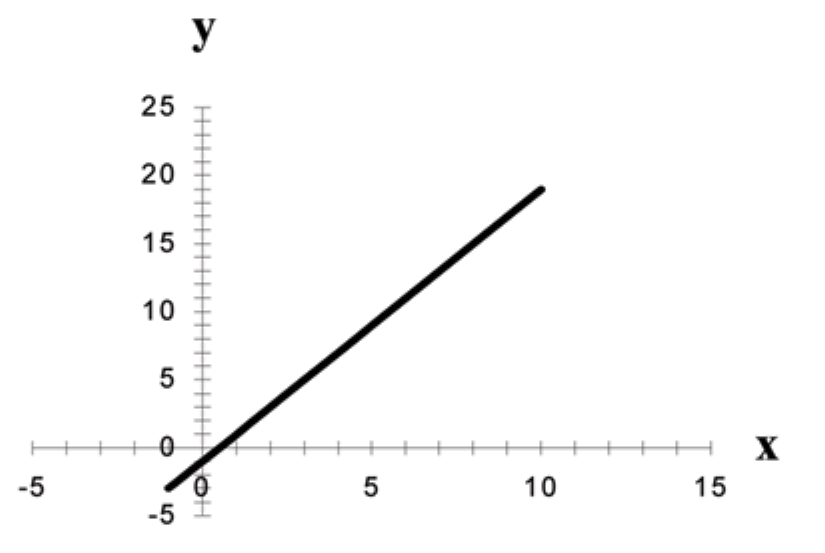
$$y=3+2\mathrm{x}$$

Equation:

$$y=-0.01+1.2\mathrm{x}$$

The graph of a linear equation of the form $y=a+\mathrm{bx}$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example:



Graph of the equation $y = -1 + 2$x.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example:
Aaron's Word Processing Service (AWPS) does word processing. Its rate is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.
Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

Solution:

Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

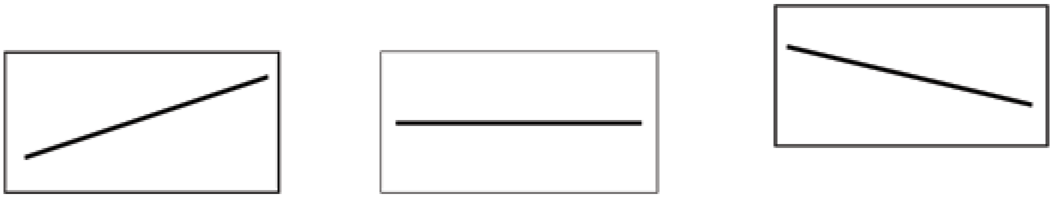
The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is:

$y = 31.50 + 32\text{x}$

Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + \mathrm{bx}$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.



If $b > 0$, the line slopes upward to the right.

If $b = 0$, the line is horizontal.

If $b < 0$, the line slopes downward to the right.

Three possible graphs of $y = a + \mathrm{bx}$.

Example:
Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of \$25 plus \$15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15\mathrm{x}$.

Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

The Regression Equation
Linear Regression and Correlation: The Regression Equation is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. Contributions from Roberta Bloom include instructions for finding and graphing the regression equation and scatterplot using the LinRegTTest on the TI-83,83+,84+ calculators.

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line**.

Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

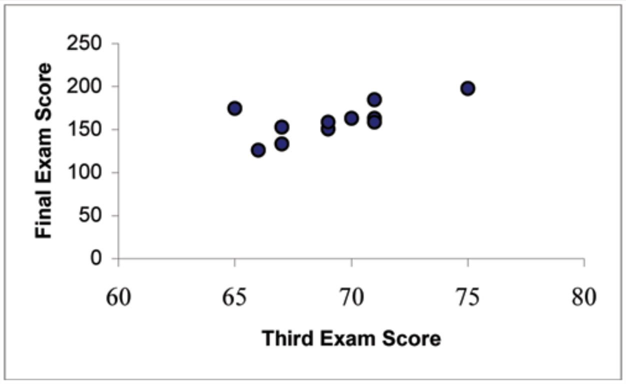
Using your equation, what is the predicted height for a pinky length of 2.5 inches?

Example:
A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

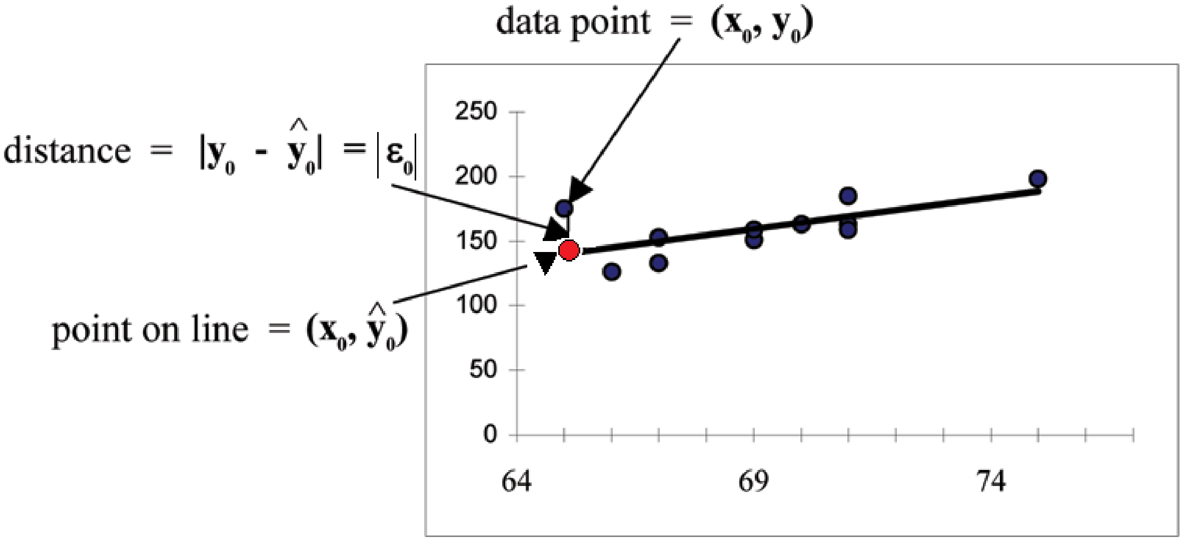
Scatter plot showing the scores on the final exam based on scores from the third exam.



The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram above, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$\left(\varepsilon_1\right)^2 + \left(\varepsilon_2\right)^2 + ... + \left(\varepsilon_{11}\right)^2 = \sum_{\mathrm{i}=1}^{11} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

Equation:

$$\hat{y} = a + \mathrm{bx}$$

where $a = \bar{y} - b \cdot \bar{x}$ and $b = \frac{\Sigma(x-\bar{x})\cdot(y-\bar{y})}{\Sigma(x-\bar{x})^2}$.

$\bar{x}$ and $\bar{y}$ are the sample means of the x values and the y values, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

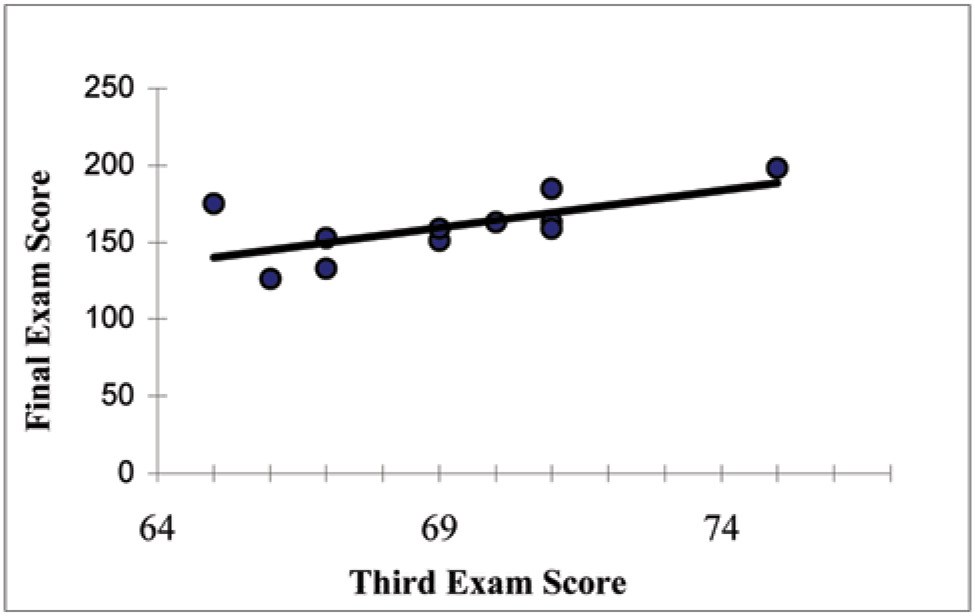
Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line** .

Note: Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:



The least squares regression line (best fit line) for the third exam/final exam example has the equation:

Equation:

$$\hat{y} = -173.51 + 4.83\text{x}$$

Note:

- Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**
- You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.
- You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

- Slope: The slope of the line is b = 4.83.

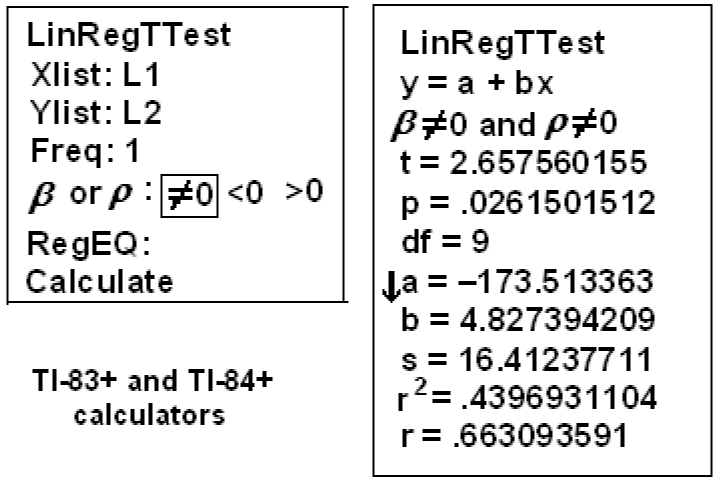
- Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)
On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
On the next line, at the prompt β or ρ, highlight "≠ 0" and press ENTER
Leave the line for "RegEq:" blank
Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen



```
LinRegTTest
Xlist: L1
Ylist: L2
Freq: 1
β or ρ : ≠0  <0  >0
RegEQ:
Calculate
```

```
LinRegTTest
y = a + bx
β≠0 and ρ≠0
t = 2.657560155
p = .0261501512
df = 9
↓a = −173.513363
b = 4.827394209
s = 16.41237711
r² = .4396931104
r = .663093591
```

TI-83+ and TI-84+ calculators

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

- The second line says y=a+bx. Scroll down to find the values a=-173.513, and b=4.8273 ; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$
- The two items at the bottom are r^2 = .43969 and r=.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

We are assuming your X data is already entered in list L1 and your Y data is in list L2
Press 2nd STATPLOT ENTER to use Plot 1
On the input screen for PLOT 1, highlight**On**and press ENTER
For TYPE: highlight the very first icon which is the scatterplot and press ENTER
Indicate Xlist: L1 and Ylist: L2
For Mark: it does not matter which symbol you highlight.
Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
To graph the best fit line, press the "Y=" key and type the equation -173.5+4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

**With contributions from Roberta Bloom

Correlation Coefficient and Coefficient of Determination
Linear Regression and Correlation: The Correlation Coefficient and Coefficient of Determination is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom. The name has been changed from Correlation Coefficient.

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y.

The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is a numerical measure of the strength of association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as
Equation:

$$r = \frac{n \cdot \mathit{\Sigma} x \cdot y - (\mathit{\Sigma} x) \cdot (\mathit{\Sigma} y)}{\sqrt{[n \cdot \mathit{\Sigma} x^2 - (\mathit{\Sigma} x)^2] \cdot [n \cdot \mathit{\Sigma} y^2 - (\mathit{\Sigma} y)^2]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.
What the VALUE of r tells us:

- The value of r is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to -1 or to +1 indicate a stronger linear relationship between x and y.
- If r=0 there is absolutely no linear relationship between x and y **(no linear correlation)**.

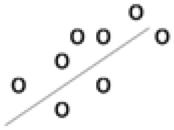
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best fit line.

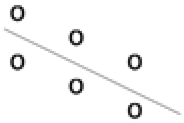
Note: Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."** For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!

Positive Correlation



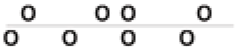
A scatter plot showing data with a positive correlation. $0 < r < 1$

Negative Correlation



A scatter plot showing data with a negative correlation. $-1 < \mathrm{r} < 0$

Zero Correlation



A scatter plot showing data with zero correlation. r =0

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

r^2 is called the coefficient of determination. r^2 is the square of the correlation coefficient , but is usually stated as a percent, rather than in decimal form. r^2 has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- 1-r^2, when expressed as a percent, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83$x
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is r^2 = 0.6631^2 = 0.4397
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best fit regression line.
- Therefore approximately 56% of the variation (1 - 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best fit regression line. (This is seen as the scattering of the points about the line.)

**With contributions from Roberta Bloom.

Glossary

Coefficient of Correlation

A measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable. The formula is:

Equation:

$$r = \frac{n\sum \mathrm{xy} - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}},$$

where n is the number of data points. The coefficient cannot be more then 1 and less then -1. The closer the coefficient is to ± 1, the stronger the evidence of a significant linear relationship between x and y.

Guessing Correlations Simulation

Begin by answering the questions, even if you have to guess. The first time you answer the questions you will not be told whether you are correct or not.

Once you have answered all the questions, answer them again using the simulation to help you. This time you will get feedback about each individual answer.

Show Simulation

This demonstration does not have questions associated with it. Use it to improve your understanding of what scatter plots of data with different values of r lool like.

General Instructions

This demonstration allows you to learn about Pearson's correlation by viewing scatter plots with different values of Pearson's r. In each case, you will have an opportunity to guess the correlation. With a little practice, you should get pretty good at it.

Step by Step Instructions

Show Questions

Click the button next to the correlation that you think is present in the scatter plot. You will find out if you were right or wrong. If you were wrong, try another value. Then click the "New data" button.
Applet failed to run. No Java plug-in was found.

Summary

After looking at a series of scatterplots, you should have a pretty good feel for what a certain value of r means.

Prediction
Linear Regression and Correlation: Prediction is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

Recall the third exam/final exam example.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best fit line for the final exam grade as a function of the grade on the third exam. We can now use the least squares regression line for prediction.

Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The exam scores **(x-values)** range from 65 to 75. **Since 73 is between the x-values 65 and 75**, substitute $x = 73$ into the equation. Then:

Equation:

$$\hat{y} = -173.51 + 4.83\left(73\right) = 179.08$$

We predict that statistic students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example:

Recall the third exam/final exam example.

Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution:

145.27

Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution:

The x values in the data are between 65 and 75. 90 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter x into the equation and calculate a y value, you should not do so!)

To really understand how unreliable the prediction can be outside of the observed x values in the data, make the substitution x = 90 into the equation.

$$\hat{y} = -173.51 + 4.83\left(90\right) = 261.19$$

The final exam score is predicted to be 261.19. The largest the final exam score can be is 200.

Note: The process of predicting inside of the observed x values in the data is called **interpolation**. The process of predicting outside of the observed x values in the data is called **extrapolation**.

**With contributions from Roberta Bloom

Linear Regression and Correlation: Summary

Bivariate Data: Each data point has two values. The form is .

Line of Best Fit or Least Squares Line (LSL):

= independent variable; = dependent variable

Residual:
Correlation Coefficient r:

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer is to 1 or -1, the closer the original points are to a straight line.
3. If is negative, the slope is negative. If is positive, the slope is positive.
4. If , then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the **SSE**, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

Practice: Linear Regression
This module provides a practice of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will evaluate bivariate data and determine if a line is an appropriate fit to the data.

Given

Below are real data for the first two decades of AIDS reporting. (*Source: Centers for Disease Control and Prevention, National Center for HIV, STD, and TB Prevention*)

Year	**# AIDS cases diagnosed**	**# AIDS deaths**
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987

1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Note: We will use the columns “year” and “# AIDS cases diagnosed” for all questions unless otherwise stated.

Graphing

Graph “year” vs. “# AIDS cases diagnosed.” **Plot the points on the graph located below in the section titled "Plot"** . Do not include pre-1981. Label both axes with words. Scale both axes.

Data

Exercise:

Problem:

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Linear Equation

Write the linear equation below, rounding to 4 decimal places:

Note: For any prediction questions, the answers are calculated using the least squares (best fit) line equation cited in the solution.

Exercise:

Problem: Calculate the following:

- **a** $a =$
- **b** $b =$

- **c** $\text{corr.} =$
- **d** $n =$(# of pairs)

Solution:

- **a** $a =$ -3,448,225
- **b** $b = 1750$
- **c** $\text{corr.} = 0.4526$
- **d** $n = 22$

Exercise:

Problem: equation: $\hat{y} =$

Solution:

$\hat{y} =$ -3,448,225 $+1750x$

Solve

Exercise:

Problem: Solve.

- **a** When $x = 1985$, $\hat{y} =$
- **b** When $x = 1990$, $\hat{y} =$

Solution:

- **a** 25,525
- **b** 34,275

Plot

Plot the 2 above points on the graph below. Then, connect the 2 points to form the regression line.

Obtain the graph on your calculator or computer.

Discussion Questions

Look at the graph above.

Exercise:

Problem: Does the line seem to fit the data? Why or why not?

Exercise:

Problem: Do you think a linear fit is best? Why or why not?

Exercise:

Problem:

Hand draw a smooth curve on the graph above that shows the flow of the data.

Exercise:

Problem:

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Exercise:

Problem:

Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Exercise:

Problem: Solve.

- **a**When $x = 1970$, $\hat{y} =$:
- **b**Why doesn’t this answer make sense?

Solution:

- **a** -725

Homework
Linear Regression and Correlation: Homework is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.
Exercise:

Problem: For each situation below, state the independent variable and the dependent variable.

- **a**A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than all other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
- **b**A study is done to determine if the weekly grocery bill changes based on the number of family members.
- **c**Insurance companies base life insurance premiums partially on the age of the applicant.
- **d**Utility bills vary according to power consumption.
- **e**A study is done to determine if a higher education reduces the crime rate in a population.

Solution:

- **a**Independent: Age; Dependent: Fatalities
- **d**Independent: Power Consumption; Dependent: Utility

Note: For any prediction questions, the answers are calculated using the least squares (best fit) line equation cited in the solution.

Exercise:

Problem:

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows (Source: *http:// http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html*):

Age	**Number of Driver Deaths per 100,000**
16-19	38
20-24	36
25-34	24
35-54	20
55-74	18

Age	Number of Driver Deaths per 100,000
75+	28

- **a**For each age group, pick the midpoint of the interval for the x value. (For the 75+ group, use 80.)
- **b**Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
- **c**Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Pick two ages and find the estimated fatality rates.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between age of a driver and driver fatality rate?
- **h**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

The average number of people in a family that received welfare for various years is given below. (Source: *House Ways and Means Committee, Health and Human Services Department*)

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

- **a**Using "year" as the independent variable and "welfare family size" as the dependent variable, make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **c**Find the correlation coefficient. Is it significant?
- **d**Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
- **e**Use the two points in (d) to plot the least squares line on your graph from (b).
- **f**Based on the above data, is there a linear relationship between the year and the average number of people in a welfare family?

- **g**Using the least squares line, estimate the welfare family sizes for 1960 and 1995. Does the least squares line give an accurate estimate for those years? Explain why or why not.
- **h**Are there any outliers in the above data?
- **i**What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b** $\hat{y} = 88.7206 - 0.0432x$
- **c**-0.8533, Yes
- **g**No
- **h**No.
- **i**2.93, Yes
- **j**slope = -0.0432. As the year increases by one, the welfare family size tends to decrease by 0.0432 people.

Exercise:

Problem:

Use the AIDS data from the practice for this section, but this time use the columns “year #” and “# new AIDS deaths in U.S.” Answer all of the questions from the practice again, using the new columns.

Exercise:

Problem:

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Source: *Microsoft Bookshelf*)

Height (in feet)	Stories
1050	57
428	28
362	26
529	40
790	60
401	22
380	38
1454	110

Height (in feet)	**Stories**
1127	100
700	46

- **a**Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
- **b**Does it appear from inspection that there is a relationship between the variables?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated heights for 32 stories and for 94 stories.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
- **h**Are there any outliers in the above data? If so, which point(s)?
- **i**What is the estimated height of a building with 6 stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
- **j**Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = 102.4287 + 11.7585x$
- **d**0.9436; yes
- **e**478.70 feet; 1207.73 feet
- **g**Yes
- **h**Yes; $(57{,}1050)$
- **i**172.98; No
- **j**11.7585 feet
- **k**slope = 11.7585. As the number of stories increases by one, the height of the building tends to increase by 11.7585 feet.

Exercise:

Problem:

Below is the life expectancy for an individual born in the United States in certain years. (Source: *National Center for Health Statistics*)

Year of Birth	**Life Expectancy**
1930	59.7

Year of Birth	**Life Expectancy**
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Draw a scatter plot of the ordered pairs.
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
- **f**Why aren't the answers to part (e) the values on the above chart that correspond to those years?
- **g**Use the two points in (e) to plot the least squares line on your graph from (b).
- **h**Based on the above data, is there a linear relationship between the year of birth and life expectancy?
- **i**Are there any outliers in the above data?
- **j**Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

The percent of female wage and salary workers who are paid hourly rates is given below for the years 1979 - 1992. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	**Percent of workers paid hourly rates**
1979	61.2
1980	60.7
1981	61.3

Year	Percent of workers paid hourly rates
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- **a**Using “year” as the independent variable and “percent” as the dependent variable, make a scatter plot of the data.
- **b**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated percents for 1991 and 1988.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
- **h**Are there any outliers in the above data?
- **i**What is the estimated percent for the year 2050? Does the least squares line give an accurate estimate for that year? Explain why or why not?
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = -266.8863 + 0.1656x$
- **d**0.9448; Yes
- **e**62.8233; 62.3265
- **h** yes; (1987, 62.7)
- **i**72.5937; No
- **j**slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Exercise:

Problem:

The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition 10, for various pages is given below.

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Draw a scatter plot of the ordered pairs.
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated maximum values for the restaurants on page 10 and on page 70.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
- **h**Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
- **i**Is the least squares line valid for page 200? Why or why not?
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

The next two questions refer to the following data: The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Exercise:

Problem:

- **a**Using “size” as the independent variable and “cost” as the dependent variable, make a scatter plot.
- **b**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**If the laundry detergent were sold in a 40 ounce size, find the estimated cost.
- **f**If the laundry detergent were sold in a 90 ounce size, find the estimated cost.
- **g**Use the two points in (e) and (f) to plot the least squares line on your graph from (a).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost? Why or why not?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = 3.5984 + 0.0371x$
- **d**0.9986; Yes
- **e**$5.08
- **f**$6.93
- **i**No
- **j**Not valid
- **k**slope = 0.0371. As the number of ounces increases by one, the cost of liquid detergent tends to increase by $0.0371 or is predicted to increase by $0.0371 (about 4 cents).

Exercise:

Problem:

- **a**Complete the above table for the cost per ounce of the different sizes.
- **b**Using “Size” as the independent variable and “Cost per ounce” as the dependent variable, make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**If the laundry detergent were sold in a 40 ounce size, find the estimated cost per ounce.
- **g**If the laundry detergent were sold in a 90 ounce size, find the estimated cost per ounce.
- **h**Use the two points in (f) and (g) to plot the least squares line on your graph from (b).
- **i**Does it appear that a line is the best way to fit the data? Why or why not?
- **j**Are there any outliers in the above data?
- **k**Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost per ounce? Why or why not?
- **l**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

According to flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	**Approximate Probate Fees and Taxes ($)**
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Based on the above, what would be the probate fees and taxes for an estate that does not have any assets?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **c**Yes
- **d** $\hat{y} = -337{,}424.6478 + 0.5463x$
- **e**0.9964; Yes
- **f**$208,875.35; $1,028,325.35
- **h**Yes
- **i**No
- **k**slope = 0.5463. As the net taxable estate increases by one dollar, the approximate probate fees and taxes tend to increase by 0.5463 dollars (about 55 cents).

Exercise:

Problem: The following are advertised sale prices of color televisions at Anderson’s.

Size (inches)	**Sale Price ($)**
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y}=a+\text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem: Below are the average heights for American boys. (Source: *Physician’s Handbook, 1990*)

Age (years)	**Height (cm)**
birth	50.8
2	83.8

Age (years)	Height (cm)
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**Find the estimated average height for a one year–old. Find the estimated average height for an eleven year–old.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Use the least squares line to estimate the average height for a sixty–two year–old man. Do you think that your answer is reasonable? Why or why not?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **c**Yes
- **d** $\hat{y} = 65.0876 + 7.0948x$
- **e**0.9761; yes
- **f**72.2 cm; 143.13 cm
- **h**Yes
- **i**No
- **j**505.0 cm; No
- **k**slope = 7.0948. As the age of an American boy increases by one year, the average height tends to increase by 7.0948 cm.

Exercise:

Problem:

The following chart gives the gold medal times for every other Summer Olympics for the women's 100 meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is the decrease in times significant?
- **f**Find the estimated gold medal time for 1932. Find the estimated time for 1984.
- **g**Why are the answers from (f) different from the chart values?
- **h**Use the two points in (f) to plot the least squares line on your graph from (b).
- **i**Does it appear that a line is the best way to fit the data? Why or why not?
- **j**Use the least squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

The next three questions use the following state information.

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado		1876	38	104,100

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Hawaii		1959	50	10,932
Iowa		1846	29	56,276
Maryland		1788	7	12,407
Missouri		1821	24	69,709
New Jersey		1787	3	8,722
Ohio		1803	17	44,828
South Carolina	13	1788	8	32,008
Utah		1896	45	84,904
Wisconsin		1848	30	65,499

Exercise:

Problem:

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Use the least squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

Solution:

- **c**No
- **d** $\hat{y} = 47.03 - 0.0216x$
- **e**-0.4280
- **f**6; 5

Exercise:

Problem:

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

- **a**Let rank be the independent variable and area be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers?
- **j**Use the least squares line to estimate the area of a new state that enters the Union. Can the least squares line be used to predict it? Why or why not?
- **k**Delete “Hawaii” and substitute “Alaska” for it. Alaska is the fortieth state with an area of 656,424 square miles.
- **l**Calculate the new least squares line.
- **m**Find the estimated area for Alabama. Is it closer to the actual area with this new least squares line or with the previous one that included Hawaii? Why do you think that’s the case?
- **n**Do you think that, in general, newer states are larger than the original states?

Exercise:

Problem:

We are interested in whether there is a relationship between the rank of a state and the year it entered the Union.

- **a**Let year be the independent variable and rank be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Why must the relationship be positive between the variables?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Let’s say a fifty-first state entered the union. Based upon the least squares line, when should that have occurred?
- **g**Using the least squares line, how many states do we currently have?
- **h**Why isn’t the least squares line a good estimator for this year?

Solution:

- **d** $\hat{y} = -480.5845 + 0.2748x$
- **e**0.9553
- **f**1934

Exercise:

Problem:

Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a union. We are interested in whether the decrease is significant. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	**Percent**
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1993	15.8
2011	11.8

- **a**Let year be the independent variable and percent be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Why will the relationship between the variables be negative?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Based on your answer to (e), do you think that the relationship can be said to be decreasing?
- **g**If the trend continues, when will there no longer be any union members? Do you think that will happen?

The next two questions refer to the following information: The data below reflects the 1991-92 Reunion Class Giving. (Source: *SUNY Albany alumni magazine*)

Class Year	**Average Gift**	**Total Giving**
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772

Class Year	Average Gift	Total Giving
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

Exercise:

Problem:

We will use the columns “class year” and “total giving” for all questions, unless otherwise stated.

- **a**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y}=a+\text{bx}$
- **c**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **d**For the class of 1930, predict the total class gift.
- **e**For the class of 1964, predict the total class gift.
- **f**For the class of 1850, predict the total class gift. Why doesn’t this value make any sense?

Solution:

- **b** $\hat{y}=-569{,}770.2796+296.0351x$
- **c**0.8302
- **d**$1577.46
- **e**$11,642.66
- **f**-$22,105.34

Exercise:

Problem:

We will use the columns “class year” and “average gift” for all questions, unless otherwise stated.

- **a**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y}=a+\text{bx}$
- **c**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **d**For the class of 1930, predict the average class gift.
- **e**For the class of 1964, predict the average class gift.
- **f**For the class of 2010, predict the average class gift. Why doesn’t this value make any sense?

Exercise:

Problem:

We are interested in exploring the relationship between the weight of a vehicle and its fuel efficiency (gasoline mileage). The data in the table show the weights, in pounds, and fuel efficiency, measured in miles per gallon, for a sample of 12 vehicles.

Weight	**Fuel Efficiency**
2715	24
2570	28
2610	29
2750	38
3000	25
3410	22
3640	20
3700	26
3880	21
3900	18
4060	18
4710	15

- **a**Graph a scatterplot of the data.
- **b**Find the correlation coefficient and determine if it is significant.
- **c**Find the equation of the best fit line.
- **d**Write the sentence that interprets the meaning of the slope of the line in the context of the data.
- **e**What percent of the variation in fuel efficiency is explained by the variation in the weight of the vehicles, using the regression line? (State your answer in a complete sentence in the context of the data.)
- **f**Accurately graph the best fit line on your scatterplot.
- **g**For the vehicle that weights 3000 pounds, find the residual (y-yhat). Does the value predicted by the line underestimate or overestimate the observed data value?
- **h**Identify any outliers, using either the graphical or numerical procedure demonstrated in the textbook.
- **i**The outlier is a hybrid car that runs on gasoline and electric technology, but all other vehicles in the sample have engines that use gasoline only. Explain why it would be appropriate to remove the

outlier from the data in this situation. Remove the outlier from the sample data. Find the new correlation coefficient, coefficient of determination, and best fit line.

- **j**Compare the correlation coefficients and coefficients of determination before and after removing the outlier, and explain in complete sentences what these numbers indicate about how the model has changed.

Solution:

- **b**r = **-**0.8, significant
- **c**yhat = 48.4-0.00725x
- **d**For every one pound increase in weight, the fuel efficiency tends to decrease (or is predicted to decrease) by 0.00725 miles per gallon. (For every one thousand pounds increase in weight, the fuel efficiency tends to decrease by 7.25 miles per gallon.)
- **e**64% of the variation in fuel efficiency is explained by the variation in weight using the regression line.
- **g**yhat=48.4-0.00725(3000)=26.65 mpg. y-yhat=25-26.65=-1.65. Because yhat=26.5 is greater than y=25, the line overestimates the observed fuel efficiency.
- **h**(2750,38) is the outlier. Be sure you know how to justify it using the requested graphical or numerical methods, not just by guessing.
- **i**yhat = 42.4-0.00578x
- **j**Without outlier, r=**-**0.885, rsquare=0.76; with outlier, r=**-**0.8, rsquare=0.64. The new linear model is a better fit, after the outlier is removed from the data, because the new correlation coefficient is farther from 0 and the new coefficient of determination is larger.

Exercise:

Problem:

The four data sets below were created by statistician Francis Anscomb. They show why it is important to examine the scatterplots for your data, in addition to finding the correlation coefficient, in order to evaluate the appropriateness of fitting a linear model.

Set 1			Set 2			Set 3			Set 4	
x	y		x	y		x	y		x	y
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.10		14	8.84		8	7.04

6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.10		4	5.39		19	12.50
12	10.84		12	9.13		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

a. For each data set, find the least squares regression line and the correlation coefficient. What did you discover about the lines and values of r?

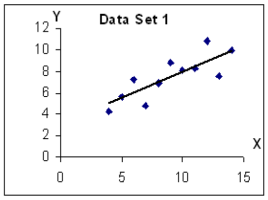
For each data set, create a scatter plot and graph the least squares regression line. Use the graphs to answer the following questions:

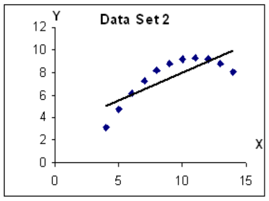
- **b**For which data set does it appear that a curve would be a more appropriate model than a line?
- **c**Which data set has an **influential point** (point close to or on the line that greatly influences the best fit line)?
- **d**Which data set has an **outlier** (obviously visible on the scatter plot with best fit line graphed)?
- **e**Which data set appears to be the most appropriate to model using the least squares regression line?

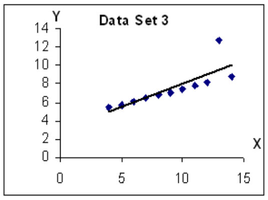
Solution:

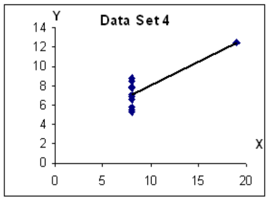
a. All four data sets have the same correlation coefficient r=0.816 and the same least squares regression line yhat=3+0.5x

b. Set 2 ; c. Set 4 ; d. Set 3 ; e. Set 1









Try these multiple choice questions

Exercise:

Problem: A correlation coefficient of -0.95 means there is a ____________ between the two variables.

- **A**Strong positive correlation
- **B**Weak negative correlation

- **C**Strong negative correlation
- **D**No Correlation

Solution:

C

Exercise:

Problem:

According to the data reported by the New York State Department of Health regarding West Nile Virus **(http://www.health.state.ny.us/nysdoh/westnile/update/update.htm)** for the years 2000-2008, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is $\hat{y} = -10.2638 + 0.0491x$. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected? $r = 0.5490$

- **A**No prediction can be made.
- **B**19.6
- **C**15
- **D**38.1

Solution:

A

The next three questions refer to the following data: (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from *www.nhc.noaa.gov/gifs/table6.gif* A major hurricane is one with a strength rating of 3, 4 or 5.

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	9	3

Exercise:

Problem:

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based upon the total number of hurricanes.

- **A** $\hat{y} = -1.67x + 0.5$
- **B** $\hat{y} = 0.5\text{x} - 1.67$
- **C** $\hat{y} = 0.94x - 1.67$
- **D** $\hat{y} = -2\text{x} + 1$

Solution:

B

Exercise:

Problem: The correlation coefficient is 0.942. Is this considered significant? Why or why not?

- **A**No, because 0.942 is greater than the critical value of 0.707
- **B**Yes, because 0.942 is greater than the critical value of 0.707
- **C**No, because 0942 is greater than the critical value of 0.811
- **D**Yes, because 0.942 is greater than the critical value of 0.811

Solution:

D

Exercise:

Problem:

The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit mainland U.S. Can the least squares line be used to make this prediction?

- **A**No, because 9 lies outside the independent variable values
- **B**Yes, because, in fact, there have been 3 major hurricanes this decade
- **C**No, because 2.83 lies outside the dependent variable values
- **D**Yes, because how else could we predict what is going to happen this decade.

Solution:

A

**Exercises 21 and 22 contributed by Roberta Bloom

Introduction
This module introduces the concept of Probability, the chance of an event occurring.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams (optional).
- Construct and interpret Tree Diagrams (optional).

Introduction

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn to solve probability problems using a systematic approach.

Optional Collaborative Classroom Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $\mathrm{P(change)}$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $\mathrm{P(bus)}$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $\mathrm{P(change)}$.
- Find $\mathrm{P(bus)}$.
- Find $\mathrm{P(change\ and\ bus)}$ Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $\mathrm{P(change|\ bus)}$ Find the probability that a randomly chosen student has change given that he/she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

Terminology
Probability: Terminology is part of the collection col10555 written by Barbara Illowsky and Susan Dean defines key terms related to Probability and has contributions from Roberta Bloom.

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

The result of an experiment is called an **outcome**. A **sample space** is a set of all possible outcomes. Three ways to represent a sample space are to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $\mathrm{S} = \{\mathrm{H}, \mathrm{T}\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $\mathrm{P}(\mathrm{A})$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between 0 and 1, inclusive** (includes 0 and 1 and all numbers between these values). $\mathrm{P}(\mathrm{A}) = 0$ means the event A can never happen. $\mathrm{P}(\mathrm{A}) = 1$ means the event A always happens. $\mathrm{P}(\mathrm{A}) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative fequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head(H) and a Tail(T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{\mathrm{HH, TH, HT, TT}\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes $\{\mathrm{HT, TH}\}$. $\mathrm{P(A)} = \frac{2}{4}$.

Suppose you roll one fair six-sided die, with the numbers {1,2,3,4,5,6} on its faces. Let event E = rolling a number that is at least 5. There are two outcomes $\{5, 6\}$. $\mathrm{P(E)} = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, 2/6 of the rolls would result in an outcome of "at least 5". You would not expect exactly 2/6. The long-term relative frequency of obtaining this result would approach the theoretical probability of 2/6 as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is the known as the **Law of Large Numbers**: as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes don't happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.) The Law of Large Numbers will be discussed again in Chapter 7.

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased** . Two math professors in Europe had their statistics students test the Belgian 1 Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos have a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later in this chapter we will learn techniques to use to work with probabilities for events that are not equally likely.

"OR" Event:
An outcome is in the event $A\text{ OR }B$ if the outcome is in A or is in B or is in both A and B. For example, let $\mathrm{A}=\{1, 2, 3, 4, 5\}$ and $\mathrm{B}=\{4, 5, 6, 7, 8\}$. $A\text{ OR }\mathrm{B}\ =\ \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

"AND" Event:
An outcome is in the event $\mathrm{A\ AND\ B}$ if the outcome is in both A and B at the same time. For example, let A and B be $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7, 8\}$, respectively. Then $\mathrm{A\ AND\ B}=\{4, 5\}$.

The **complement** of event A is denoted $\mathrm{A'}$ (read "A prime"). $\mathrm{A'}$ consists of all outcomes that are **NOT** in A. Notice that $\mathrm{P(A)+P(A')=1}$. For example, let $\mathrm{S}=\{1, 2, 3, 4, 5, 6\}$ and let $\mathrm{A}=\{1, 2, 3, 4\}$. Then, $\mathrm{A'}=\{5, 6\}$. $\mathrm{P(A)}=\frac{4}{6}$, $\mathrm{P(A')}=\frac{2}{6}$, and $\mathrm{P(A)+P(A')}=\frac{4}{6}+\frac{2}{6}=1$

The **conditional probability** of A given B is written $\mathrm{P(A|B)}$. $\mathrm{P(A|B)}$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space**. We calculate the probability of A from the reduced sample space B. The formula to calculate $\mathrm{P(A|B)}$ is

$$\mathrm{P(A|B)}=\frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}}$$

where $\mathrm{P(B)}$ is greater than 0.

For example, suppose we toss one fair, six-sided die. The sample space $\mathrm{S}=\{1, 2, 3, 4, 5, 6\}$. Let A = face is 2 or 3 and B = face is even (2, 4, 6). To calculate $\mathrm{P(A|B)}$, we count the number of outcomes 2 or 3 in the sample space $\mathrm{B} = \{2,\ 4,\ 6\}$. Then we divide that by the number of outcomes in B (and not S).

We get the same result by using the formula. Remember that S has 6 outcomes.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A\ and\ B)}}{\mathrm{P(B)}} = \frac{\text{(the number of outcomes that are 2 or 3 and even in S) / 6}}{\text{(the number of outcomes that are even in S) / 6}} = \frac{1/6}{3/6} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Exercise:

Problem:

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts (a) through (j) below. (Note that you can't find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

- **a** The probability that a student does not have long hair.
- **b** The probability that a student is male or has short hair.
- **c** The probability that a student is a female and has long hair.
- **d** The probability that a student is male, given that the student has long hair.

- **e** The probability that a student has long hair, given that the student is male.
- **f** Of all the female students, the probability that a student has short hair.
- **g** Of all students with long hair, the probability that a student is female.
- **h** The probability that a student is female or has long hair.
- **i** The probability that a randomly selected student is a male student with short hair.
- **j** The probability that a student is female.

Solution:

- **a** P(L')=P(S)
- **b** P(M or S)
- **c** P(F and L)
- **d** P(M|L)
- **e** P(L|M)
- **f** P(S|F)
- **g** P(F|L)
- **h** P(F or L)
- **i** P(M and S)
- **j** P(F)

**With contributions from Roberta Bloom

Glossary

Conditional Probability

The likelihood that an event will occur given that another event has already occurred.

Equally Likely

Each outcome of an experiment has the same probability.

Experiment

A planned activity carried out under controlled conditions.

Event

A subset in the set of all outcomes of an experiment. The set of all outcomes of an experiment is called a **sample space** and denoted usually by S. An event is any arbitrary subset in **S**. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, etc. Standard notations for events are capital letters such as A, B, C, etc.

Outcome (observation)

A particular result of an experiment.

Probability

A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1;$.
- If A and B are any two mutually exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$.
- $P(S) = 1$.

Sample Space

The set of all possible outcomes of an experiment.

Independent and Mutually Exclusive Events
Probability: Independent and Mutually Exclusive Events is part of the collection col10555 written by Barbara Illowsky and Susan Dean and explains the concept of independent events, where the probability of event A does not have any effect on the probability of event B, and mutually exclusive events, where events A and B cannot occur at the same time. The module has contributions from Roberta Bloom.

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- $\mathrm{P(A|B) = P(A)}$
- $\mathrm{P(B|A) = P(B)}$
- $\mathrm{P(A\ AND\ B) = P(A) \cdot P(B)}$

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:**: When sampling is done without replacement, then each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of

the first pick. The events are considered to be dependent or not independent.

If it is not known whether A and B are independent or dependent, **assume they are dependent until you can show otherwise**.

Mutually Exclusive Events

A and B are **mutually exclusive** events if they cannot occur at the same time. This means that A and B do not share any outcomes and $\mathrm{P(A\ AND\ B)} = 0$.

For example, suppose the sample space $\mathrm{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $\mathrm{A} = \{1, 2, 3, 4, 5\}$, $\mathrm{B} = \{4, 5, 6, 7, 8\}$, and $\mathrm{C} = \{7, 9\}$. $\mathrm{A\ AND\ B} = \{4, 5\}$. $\mathrm{P(A\ AND\ B)} = \frac{2}{10}$ and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $\mathrm{P(A\ AND\ C)} = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**.

The following examples illustrate these definitions and terms.

Example:
Flip two fair coins. (This is an experiment.)
The sample space is $\{\mathrm{HH, HT, TH, TT}\}$ where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means 0 or 1 tail.) Then A can be written as $\{\mathrm{HH, HT, TH}\}$. The outcome HH shows 0 tails. HT and TH each show 1 tail.

- Let B = the event of getting all tails. B can be written as $\{\mathrm{TT}\}$. B is the **complement** of A. So, $\mathrm{B} = \mathrm{A'}$. Also, $\mathrm{P(A)} + \mathrm{P(B)} = \mathrm{P(A)} + \mathrm{P(A')} = 1$.
- The probabilities for A and for B are $\mathrm{P(A)} = \frac{3}{4}$ and $\mathrm{P(B)} = \frac{1}{4}$.
- Let C = the event of getting all heads. $\mathrm{C} = \{\mathrm{HH}\}$. Since $\mathrm{B} = \{\mathrm{TT}\}$, $\mathrm{P(B\ AND\ C)} = 0$. B and C are mutually exclusive. (B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $D = \{\mathrm{TT}\}$. $\mathrm{P(D)} = \frac{1}{4}$.
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $E = \{\mathrm{HT}, \mathrm{HH}\}$. $\mathrm{P(E)} = \frac{2}{4}$.
- Find the probability of getting **at least one** (1 or 2) tail in two flips. Let F = event of getting at least one tail in two flips. $F = \{\mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. $\mathrm{P(F)} = \frac{3}{4}$

Example:

Roll one fair 6-sided die. The sample space is $\{1, 2, 3, 4, 5, 6\}$. Let event A = a face is odd. Then $\mathrm{A} = \{1, 3, 5\}$. Let event B = a face is even. Then $\mathrm{B} = \{2, 4, 6\}$.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $\mathrm{P(A)} + \mathrm{P(B)} = \mathrm{P(A)} + \mathrm{P(A')} = 1$. Also, $\mathrm{P(A)} = \frac{3}{6}$ and $\mathrm{P(B)} = \frac{3}{6}$
- Let event C = odd faces larger than 2. Then $C = \{3, 5\}$. Let event D = all even faces smaller than 5. Then $D = \{2, 4\}$. $\mathrm{P(C\ and\ D)} = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than 5. $E = \{1, 2, 3, 4\}$.
  Exercise:

Problem:

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Solution:

No. C = $\{3, 5\}$ and E = $\{1, 2, 3, 4\}$. $\mathrm{P(C\ AND\ E)} = \frac{1}{6}$. To be mutually exclusive, $\mathrm{P(C\ AND\ E)}$ must be 0.

- Find $\mathrm{P(C|A)}$. This is a conditional. Recall that the event C is $\{3, 5\}$ and event A is $\{1, 3, 5\}$. To find $\mathrm{P(C|A)}$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space $\{1, 2, 3, 4, 5, 6\}$ to $\{1, 3, 5\}$. So, $\mathrm{P(C|A)} = \frac{2}{3}$

Example:

Let event G = taking a math class. Let event H = taking a science class. Then, $\mathrm{G\ AND\ H}$ = taking a math class and a science class. Suppose $\mathrm{P(G)} = 0.6$, $\mathrm{P(H)} = 0.5$, and $\mathrm{P(G\ AND\ H)} = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $\mathrm{P(G|H)} = \mathrm{P(G)}$
- $\mathrm{P(H|G)} = \mathrm{P(H)}$
- $\mathrm{P(G\ AND\ H)} = \mathrm{P(G)} \cdot \mathrm{P(H)}$

Note:The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

Exercise:

Problem: Show that $\mathrm{P(G|H) = P(G)}$.

Solution:

$$\mathrm{P(G|H)} = \frac{\mathrm{P(G\ AND\ H)}}{\mathrm{P(H)}} = \frac{0.3}{0.5} = 0.6 = \mathrm{P(G)}$$

Exercise:

Problem: Show $\mathrm{P(G\ AND\ H) = P(G) \cdot P(H)}$.

Solution:

$$\mathrm{P(G) \cdot P(H)} \ = \ 0.6 \cdot 0.5 \ = \ 0.3 \ = \ \mathrm{P(G\ AND\ H)}$$

Since G and H are independent, then, knowing that a person is taking a science class does not change the chance that he/she is taking math. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he/she is taking math. For practice, show that $\mathrm{P(H|G) = P(H)}$ to show that G and H are independent events.

Example:

In a box there are 3 red cards and 5 blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.

The sample space $S \ = \ \mathrm{R1, R2, R3, B1, B2, B3, B4, B5}$. S has 8 outcomes.

- $\mathrm{P(R)} = \frac{3}{8}$. $\mathrm{P(B)} = \frac{5}{8}$. $\mathrm{P(R\ AND\ B)} = 0$. (You cannot draw one card that is both red and blue.)

- $\mathrm{P(E)} = \frac{3}{8}$. (There are 3 even-numbered cards, $\mathrm{R2}$, $\mathrm{B2}$, and $\mathrm{B4}$.)
- $\mathrm{P(E|B)} = \frac{2}{5}$. (There are 5 blue cards: $\mathrm{B1}$, $\mathrm{B2}$, $\mathrm{B3}$, $\mathrm{B4}$, and $\mathrm{B5}$. Out of the blue cards, there are 2 even cards: $\mathrm{B2}$ and $\mathrm{B4}$.)
- $\mathrm{P(B|E)} = \frac{2}{3}$. (There are 3 even-numbered cards: $\mathrm{R2}$, $\mathrm{B2}$, and $\mathrm{B4}$. Out of the even-numbered cards, 2 are blue: $\mathrm{B2}$ and $\mathrm{B4}$.)
- The events R and B are mutually exclusive because $\mathrm{P(R\ AND\ B)} = 0$.
- Let G = card with a number greater than 3. $G = \{\mathrm{B4}, \mathrm{B5}\}$. $\mathrm{P(G)} = \frac{2}{8}$. Let H = blue card numbered between 1 and 4, inclusive. $H = \{\mathrm{B1}, \mathrm{B2}, \mathrm{B3}, \mathrm{B4}\}$. $\mathrm{P(G|H)} = \frac{1}{4}$. (The only card in H that has a number greater than 3 is $\mathrm{B4}$.) Since $\frac{2}{8} = \frac{1}{4}$, $\mathrm{P(G)} = \mathrm{P(G|H)}$ which means that G and H are independent.

Example:

In a particular college class, 60% of the students are female. 50 % of all students in the class have long hair. 45% of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that the student is female. Let L be the event that the student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $\mathrm{P(F\ )} = 0.60$; $\mathrm{P(L\ )} = 0.50$
- $\mathrm{P(F\ AND\ L)} = 0.45$
- $\mathrm{P(L|F)} = 0.75$

Note:The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know P(F|L) yet, so you can not use the second condition.

Solution 1
Check whether P(F and L) = P(F)P(L): We are given that P(F and L) = 0.45 ; but P(F)P(L) = (0.60)(0.50)= 0.30 The events of being female and having long hair are not independent because P(F and L) does not equal P(F)P(L).
Solution 2
check whether P(L|F) equals P(L): We are given that P(L|F) = 0.75 but P(L) = 0.50; they are not equal. The events of being female and having long hair are not independent.
Interpretation of Results
The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

**Example 5 contributed by Roberta Bloom

Glossary

Independent Events
: The occurrence of one event has no effect on the probability of the occurrence of any other event. Events A and B are independent if one of the following is true: (1). $P(A|B)=P(A)$; (2) $P(B|A)=P(B)$; (3) $P(A \text{ and } B)=P(A)P(B)$.

Mutually Exclusive
: An observation cannot fall into more than one class (category). Being in more than one category prevents being in a mutually exclusive category.

Two Basic Rules of Probability
This module introduces the multiplication and addition rules used when calculating probabilities.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then:
$\mathrm{P(A\ AND\ B) = P(B) \cdot P(A|B)}$.

This rule may also be written as : $\mathrm{P(A|B)} = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}}$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are **independent**, then $\mathrm{P(A|B) = P(A)}$. Then
$\mathrm{P(A\ AND\ B) = P(A|B)\ P(B)}$ becomes $\mathrm{P(A\ AND\ B) = P(A)\ P(B)}$.

The Addition Rule

If A and B are defined on a sample space, then:
$\mathrm{P(A\ OR\ B) = P(A) + P(B) - P(A\ AND\ B)}$.

If A and B are **mutually exclusive**, then $\mathrm{P(A\ AND\ B) = 0}$. Then
$\mathrm{P(A\ OR\ B) = P(A) + P(B) - P(A\ AND\ B)}$ becomes
$\mathrm{P(A\ OR\ B) = P(A) + P(B)}$.

Example:
Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is $\mathrm{P(A) = 0.6}$ and the probability that he chooses B is $\mathrm{P(B) = 0.35}$.
- $\mathrm{P(A\ and\ B) = 0}$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $\mathrm{P(A\ OR\ B) = P(A) + P(B) = 0.6 + 0.35 = 0.95}$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example:

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game.
A = the event Carlos is successful on his first attempt. $\mathrm{P(A)} = 0.65$. B = the event Carlos is successful on his second attempt. $\mathrm{P(B)} = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

Exercise:

Problem: What is the probability that he makes both goals?

Solution:

The problem is asking you to find $\mathrm{P(A\ AND\ B)} = \mathrm{P(B\ AND\ A)}$. Since $\mathrm{P(B|A)} = 0.90$:

Equation:

$$\mathrm{P(B\ AND\ A)} = \mathrm{P(B|A)\ P(A)}\ =\ 0.90{*}0.65 = 0.585$$

Carlos makes the first and second goals with probability 0.585.

Exercise:

Problem:

What is the probability that Carlos makes either the first goal or the second goal?

Solution:

The problem is asking you to find $\mathrm{P(A\ OR\ B)}$.

Equation:

$$\mathrm{P(A\ OR\ B)} = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A\ AND\ B)} = 0.65 + 0.65 - 0.585 = 0.715$$

Carlos makes either the first goal or the second goal with probability 0.715.

Exercise:

Problem: Are A and B independent?

Solution:

No, they are not, because $\mathrm{P(B\ AND\ A)}\ =\ 0.585$.

Equation:

$$\mathrm{P(B)} \ \cdot \ \mathrm{P(A)} \ = \ (0.65) \ \cdot \ (0.65) \ = \ 0.423$$

Equation:

$$0.423 \ \neq \ 0.585 \ = \ \mathrm{P(B \ AND \ A)}$$

So, $\mathrm{P(B\ AND\ A)}$ is **not** equal to $\mathrm{P(B)} \cdot \mathrm{P(A)}$.

Exercise:

Problem: Are A and B mutually exclusive?

Solution:

No, they are not because $\mathrm{P(A\ and\ B)}$ = 0.585.

To be mutually exclusive, $\mathrm{P(A\ AND\ B)}$ must equal 0.

Example:

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice 4 times a week. **Thirty** of the intermediate swimmers practice 4 times a week. **Ten** of the novice swimmers practice 4 times a week. Suppose one member of the swim team is randomly chosen. Answer the questions (Verify the answers):

Exercise:

Problem: What is the probability that the member is a novice swimmer?

Solution:

$\frac{28}{150}$

Exercise:

Problem: What is the probability that the member practices 4 times a week?

Solution:

$\frac{80}{150}$

Exercise:

Problem:

What is the probability that the member is an advanced swimmer and practices 4 times a week?

Solution:

$\frac{40}{150}$

Exercise:

Problem:

What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

Solution:

$\text{P}(\text{advanced AND intermediate}) \; = \; 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Exercise:

Problem:

Are being a novice swimmer and practicing 4 times a week independent events? Why or why not?

Solution:

No, these are not independent events.

Equation:

$$\text{P}(\text{novice AND practices 4 times per week}) = 0.0667$$

Equation:

$$\text{P}(\text{novice}) \cdot \text{P}(\text{practices 4 times per week}) = 0.0996$$

Equation:

$$0.0667 \neq 0.0996$$

Example:

Studies show that, if she lives to be 90, about 1 woman in 7 (approximately 14.3%) will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

Exercise:

Problem:

What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

Solution:

$\mathrm{P(B)} = 0.143$; $\mathrm{P(N)} = 0.85$

Exercise:

Problem:

Given that the woman has breast cancer, what is the probability that she tests negative?

Solution:

$\mathrm{P(N|B)} = 0.02$

Exercise:

Problem:

What is the probability that the woman has breast cancer AND tests negative?

Solution:

$\mathrm{P(B\ AND\ N)} = \mathrm{P(B)} \cdot \mathrm{P(N|B)} = (0.143)\ \cdot\ (0.02) = 0.0029$

Exercise:

Problem:

What is the probability that the woman has breast cancer or tests negative?

Solution:

$\mathrm{P(B\ OR\ N)} = \mathrm{P(B)} + \mathrm{P(N)} - \mathrm{P(B\ AND\ N)} = 0.143 + 0.85 - 0.0029 = 0.9901$

Exercise:

Problem: Are having breast cancer and testing negative independent events?

Solution:

No. $\mathrm{P(N) = 0.85}$; $\mathrm{P(N|B) = 0.02}$. So, $\mathrm{P(N|B)}$ does not equal $\mathrm{P(N)}$

Exercise:

Problem: Are having breast cancer and testing negative mutually exclusive?

Solution:

No. $\mathrm{P(B\ AND\ N) = 0.0029}$. For B and N to be mutually exclusive, $\mathrm{P(B\ AND\ N)}$ must be 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of any other event. Events A and B are independent if one of the following is true: (1). $P(A|B) = P(A)$; (2) $P(B|A) = P(B)$; (3) $P(A \text{ and } B) = P(A)P(B)$.

Mutually Exclusive

An observation cannot fall into more than one class (category). Being in more than one category prevents being in a mutually exclusive category.

Sample Space

The set of all possible outcomes of an experiment.

Contingency Tables
This module introduces the contingency table as a way of determining conditional probabilities.

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner. Contingincy tables provide a way of portraying data that can facilitate calculating probabilities.

Example:
Suppose a study of speeding violations and drivers who use car phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Car phone user	25	280	305
Not a car phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that $305 + 450 = 755$ and $70 + 685 = 755$. Calculate the following probabilities using the table

Exercise:

Problem: $\text{P(person is a car phone user)} =$

Solution:

$$\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$$

Exercise:

Problem: $\mathrm{P}(\text{person had no violation in the last year}) =$

Solution:

$\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

Exercise:

Problem:

$\mathrm{P}(\text{person had no violation in the last year AND was a car phone user}) =$

Solution:

$\frac{280}{755}$

Exercise:

Problem:

$\mathrm{P}(\text{person is a car phone user OR person had no violation in the last year}) =$

Solution:

$\left(\frac{305}{755} + \frac{685}{755}\right)$-$\frac{280}{755} = \frac{710}{755}$

Exercise:

Problem:

$\mathrm{P}(\text{person is a car phone user GIVEN person had a violation in the last year}) =$

Solution:

$\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

Exercise:

Problem:

$\mathrm{P}(\text{person had no violation last year GIVEN person was not a car phone user}) =$

Solution:

$\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Example:
The following table shows a random sample of 100 hikers and the areas of hiking preferred:

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Hiking Area Preference

Exercise:

Problem: Complete the table.

Solution:

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Hiking Area Preference

Exercise:

Problem:

Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

- **a**$\mathrm{P(F\ AND\ C)}$ =
- **b**$\mathrm{P(F)} \cdot \mathrm{P(C)}$ =

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Solution:

- **a** $\mathrm{P(F\ AND\ C)} = \frac{18}{100} = 0.18$
- **b** $\mathrm{P(F)} \cdot \mathrm{P(C)} = \frac{45}{100} \cdot \frac{34}{100} = 0.45 \cdot 0.34 = 0.153$

$\mathrm{P(F\ AND\ C)} \neq \mathrm{P(F)} \cdot \mathrm{P(C)}$, so the events F and C are not independent.

Exercise:

Problem:

Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male and let L = prefers hiking near lakes and streams.

- **a**What word tells you this is a conditional?
- **b**Fill in the blanks and calculate the probability: $\mathrm{P}(\_\_\_|\_\_\_) = \_\_\_$.
- **c**Is the sample space for this problem all 100 hikers? If not, what is it?

Solution:

- **a**The word 'given' tells you that this is a conditional.
- **b**$\mathrm{P(M|L)} = \frac{25}{41}$
- **c**No, the sample space for this problem is 41.

Exercise:

Problem:

Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female and let P = prefers mountain peaks.

- **a**$\text{P(F)} =$
- **b**$\text{P(P)} =$
- **c**$\text{P(F AND P)} =$
- **d**Therefore, $\text{P(F OR P)} =$

Solution:

- **a**$\text{P(F)} = \frac{45}{100}$
- **b**$\text{P(P)} = \frac{25}{100}$
- **c**$\text{P(F AND P)} = \frac{11}{100}$
- **d**$\text{P(F OR P)} = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Example:

Muddy Mouse lives in a cage with 3 doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____

Caught or Not	Door One	Door Two	Door Three	Total
Total	____	____	____	1

Door Choice

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Caught).
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Not Caught).

Verify the remaining entries.

Exercise:

Problem:

Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution:

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Door Choice

Exercise:

Problem: What is the probability that Alissa does not catch Muddy?

Solution:

$\frac{41}{60}$

Exercise:

Problem:

What is the probability that Muddy chooses Door One **OR** Door Two given that Muddy is caught by Alissa?

Solution:

$\frac{9}{19}$

Note: You could also do this problem by using a probability tree. See the Tree Diagrams (Optional) section of this chapter for examples.

Glossary

Contingency Table

The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

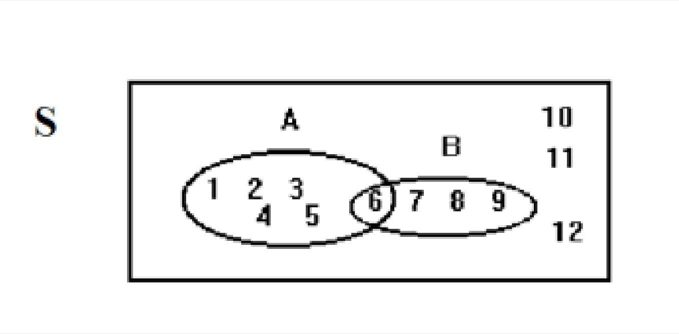
Venn Diagrams

This module introduces Venn diagrams as a method for solving some probability problems. This module is included in the Elementary Statistics textbook/collection as an optional lesson.

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example:

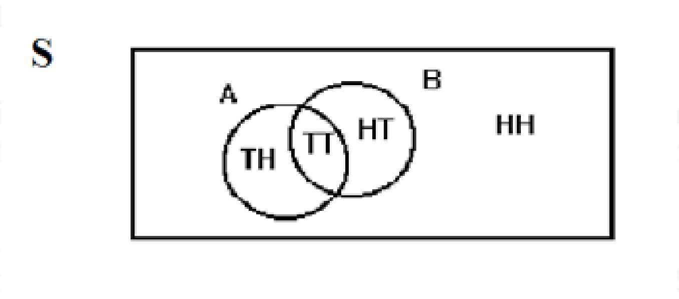
Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event $A = \{1, 2, 3, 4, 5, 6\}$ and event $B = \{6, 7, 8, 9\}$. Then $\text{A AND B} = \{6\}$ and $\text{A OR B} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The Venn diagram is as follows:



Example:

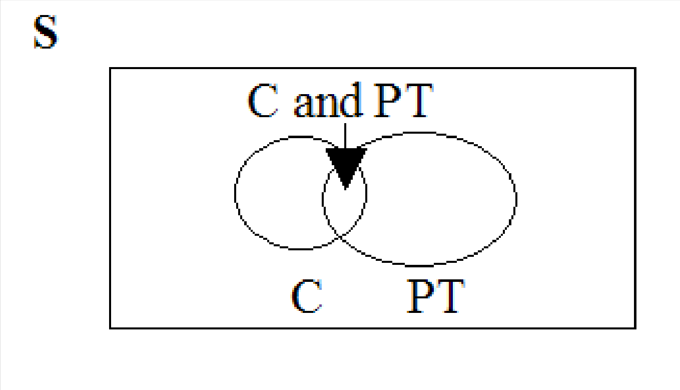
Flip 2 fair coins. Let A = tails on the first coin. Let B = tails on the second coin. Then $A = \{\text{TT}, \text{TH}\}$ and $B = \{\text{TT}, \text{HT}\}$. Therefore, $\text{A AND B} = \{\text{TT}\}$. $\text{A OR B} = \{\text{TH}, \text{TT}, \text{HT}\}$.

The sample space when you flip two fair coins is $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. The outcome HH is in neither A nor B. The Venn diagram is as follows:



Example:

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.



If a student is selected at random find

- The probability that the student belongs to a club. $\mathrm{P(C)} = 0.40$.
- The probability that the student works part time. $\mathrm{P(PT)} = 0.50$.
- The probability that the student belongs to a club AND works part time. $\mathrm{P(C\ AND\ PT)} = 0.05$.
- The probability that the student belongs to a club **given** that the student works part time.
  Equation:

$$\mathrm{P(C|PT)} = \frac{\mathrm{P(C\ AND\ PT)}}{\mathrm{P(PT)}} = \frac{0.05}{0.50} = 0.1$$

- The probability that the student belongs to a club **OR** works part time.
  Equation:

$$\mathrm{P(C\ OR\ PT)} = \mathrm{P(C)} + \mathrm{P(PT)} - \mathrm{P(C\ AND\ PT)} = 0.40 + 0.50 - 0.05 = 0.85$$

Glossary

Venn Diagram

The visual representation of a sample space and events in the form of circles or ovals showing their intersections.

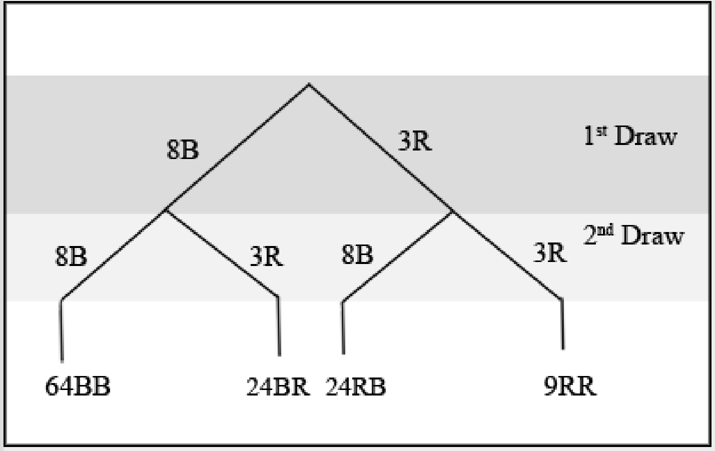
Tree Diagrams

This module introduces tree diagrams as a method for making some probability problems easier to solve. This module is included in the Elementary Statistics textbook/collection as an optional lesson.

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example:

In an urn, there are 11 balls. Three balls are red (R) and 8 balls are blue (B). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



$$\text{Total} = 64 + 24 + 24 + 9 = 121$$

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the 9 RR outcomes can be written as:

R1R1 R1R2 R1R3 R2R1 R2R2 R2R3 R3R1 R3R2 R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, and with replacement. There are $11 \cdot 11 = 121$ outcomes, the size of the **sample space**.

Exercise:

Problem: List the 24 BR outcomes: B1R1, B1R2, B1R3, ...

Solution:

B1R1 B1R2 B1R3 B2R1 B2R2 B2R3 B3R1 B3R2 B3R3 B4R1 B4R2 B4R3 B5R1 B5R2 B5R3 B6R1 B6R2 B6R3 B7R1 B7R2 B7R3 B8R1 B8R2 B8R3

Exercise:

Problem: Using the tree diagram, calculate $\mathrm{P(RR)}$.

Solution:

$\mathrm{P(RR)} = \frac{3}{11} \cdot \frac{3}{11} = \frac{9}{121}$

Exercise:

Problem: Using the tree diagram, calculate $\mathrm{P(RB\ OR\ BR)}$.

Solution:

$\mathrm{P(RB\ OR\ BR)} = \frac{3}{11} \cdot \frac{8}{11} + \frac{8}{11} \cdot \frac{3}{11} = \frac{48}{121}$

Exercise:

Problem:

Using the tree diagram, calculate $\mathrm{P(R\ on\ 1st\ draw\ AND\ B\ on\ 2nd\ draw)}$.

Solution:

$\mathrm{P(R\ on\ 1st\ draw\ AND\ B\ on\ 2nd\ draw)} = \mathrm{P(RB)} = \frac{3}{11} \cdot \frac{8}{11} = \frac{24}{121}$

Exercise:

Problem:

Using the tree diagram, calculate $\mathrm{P(R\ on\ 2nd\ draw\ given\ B\ on\ 1st\ draw)}$.

Solution:

$\mathrm{P(R\ on\ 2nd\ draw\ given\ B\ on\ 1st\ draw)} = \mathrm{P(R\ on\ 2nd} \mid \mathrm{B\ on\ 1st)} = \frac{24}{88} = \frac{3}{11}$

This problem is a conditional. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are $24 + 64 = 88$ possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

Exercise:

Problem: Using the tree diagram, calculate $\mathrm{P(BB)}$.

Solution:

$\mathrm{P(BB)} = \frac{64}{121}$

Exercise:

Problem:

Using the tree diagram, calculate
$\mathrm{P(B \text{ on the 2nd draw given R on the first draw})}$.

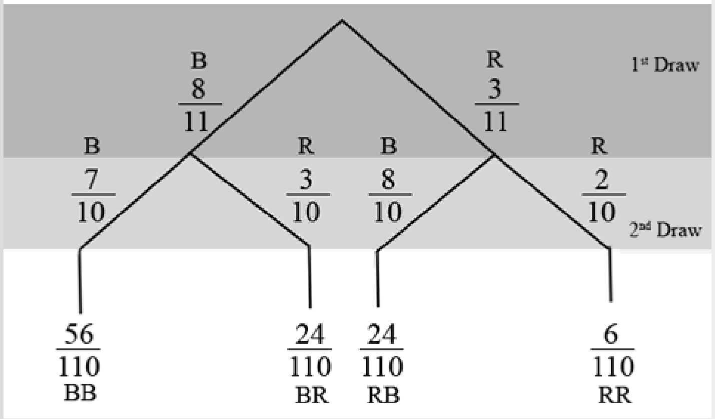
Solution:

$\mathrm{P(B \text{ on 2nd draw} \mid \text{R on 1st draw})} = \frac{8}{11}$

There are $9 + 24$ outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then $9 + 24 = 33$. Twenty-four of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Example:

An urn has 3 red marbles and 8 blue marbles in it. Draw two marbles, one at a time, this time without replacement from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second ball. Below is a tree diagram. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\frac{3}{11} \cdot \frac{2}{10} = \frac{6}{110}$.



$$\text{Total} = \frac{56+24+24+6}{110} = \frac{110}{110} = 1$$

Note: If you draw a red on the first draw from the 3 red possibilities, there are 2 red left to draw on the second draw. You do not put back or replace the first ball after you have drawn it. You draw **without replacement**, so that on the second draw there are 10 marbles left in the urn.

Calculate the following probabilities using the tree diagram.

Exercise:

Problem: P(RR) =

Solution:

$$\text{P(RR)} = \frac{3}{11} \cdot \frac{2}{10} = \frac{6}{110}$$

Exercise:

Problem: Fill in the blanks:

$$\text{P(RB OR BR)} = \frac{3}{11} \cdot \frac{8}{10} + (\_\_\_)(\_\_\_) = \frac{48}{110}$$

Solution:

$\text{P(RB or BR)} = \frac{3}{11} \cdot \frac{8}{10} + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

Exercise:

Problem: P(R on 2d | B on 1st) =

Solution:

$\text{P(R on 2d | B on 1st)} = \frac{3}{10}$

Exercise:

Problem: Fill in the blanks:

$\text{P(R on 1st and B on 2nd)} = \text{P(RB)} = (\_\_\_)(\_\_\_) = \frac{24}{110}$

Solution:

$\text{P(R on 1st and B on 2nd)} = \text{P(RB)} = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$

Exercise:

Problem: P(BB) =

Solution:

$\text{P(BB)} = \frac{8}{11} \cdot \frac{7}{10}$

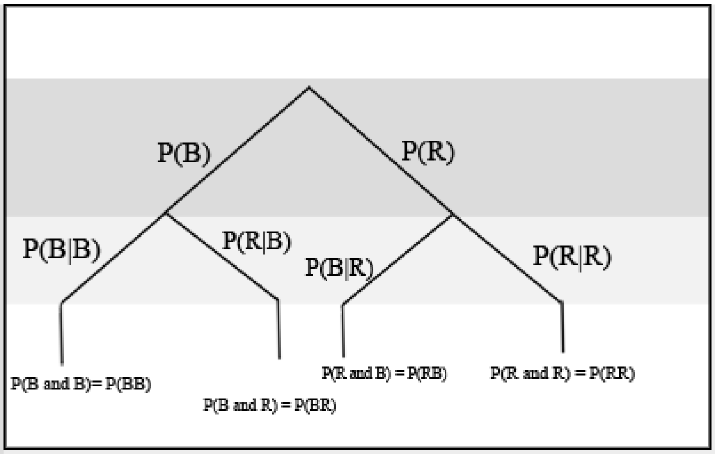
Exercise:

Problem: P(B on 2nd | R on 1st) =

Solution:

There are $6 + 24$ outcomes that have R on the first draw (6 RR and 24 RB). The 6 and the 24 are frequencies. They are also the numerators of the fractions $\frac{6}{110}$ and $\frac{24}{110}$. The sample space is no longer 110 but $6 + 24 = 30$. Twenty-four of the 30 outcomes have B on the second draw. The probability is then $\frac{24}{30}$. Did you get this answer?

If we are using probabilities, we can label the tree in the following general way.



- $\mathrm{P(R|R)}$ here means $\mathrm{P(R\ on\ 2nd\ |\ R\ on\ 1st)}$
- $\mathrm{P(B|R)}$ here means $\mathrm{P(B\ on\ 2nd\ |\ R\ on\ 1st)}$
- $\mathrm{P(R|B)}$ here means $\mathrm{P(R\ on\ 2nd\ |\ B\ on\ 1st)}$
- $\mathrm{P(B|B)}$ here means $\mathrm{P(B\ on\ 2nd\ |\ B\ on\ 1st)}$

Glossary

Sample Space

The set of all possible outcomes of an experiment.

Tree Diagram

The useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes simultaneously with associated probabilities (frequencies, relative frequencies).

Summary of Formulas

This module provides a review of the probability formulas, including the definitions of independent, complementary, and mutually exclusive events as well as the addition and multiplication rules.

Formula

Complement

If A and A' are complements then $\text{P(A)} + \text{P(A')} = 1$

Formula

Addition Rule

$\text{P(A OR B)} = \text{P(A)} + \text{P(B)} - \text{P(A AND B)}$

Formula

Mutually Exclusive

If A and B are mutually exclusive then $\text{P(A AND B)} = 0$; so $\text{P(A OR B)} = \text{P(A)} + \text{P(B)}$.

Formula

Multiplication Rule

- $\text{P(A AND B)} = \text{P(B)P(A|B)}$
- $\text{P(A AND B)} = \text{P(A)P(B|A)}$

Formula

Independence

If A and B are independent then:

- $\text{P(A|B)} = \text{P(A)}$
- $\text{P(B|A)} = \text{P(B)}$
- $\text{P(A AND B)} = \text{P(A)P(B)}$

Practice 1: Contingency Tables
This module provides the opportunity for students to apply what they've learned about probability to solve a series of problems given a set of data. Students will practice constructing and interpreting contingency tables.

Student Learning Outcomes

- The student will construct and interpret contingency tables.

Given

An article in the *New England Journal of Medicine* , reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most 10 cigarettes per day, there were 9886 African Americans, 2745 Native Hawaiians, 12,831 Latinos, 8378 Japanese Americans, and 7650 Whites. Of the people smoking 11-20 cigarettes per day, there were 6514 African Americans, 3062 Native Hawaiians, 4932 Latinos, 10,680 Japanese Americans, and 9877 Whites. Of the people smoking 21-30 cigarettes per day, there were 1671 African Americans, 1419 Native Hawaiians, 1406 Latinos, 4715 Japanese Americans, and 6062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2305 Japanese Americans, and 3970 Whites. (*(Source: http://www.nejm.org/doi/full/10.1056/NEJMoa033250)*)

Complete the Table

Complete the table below using the data provided.

Smoking Level	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
1-10						
11-20						
21-30						
31+						
TOTALS						

Smoking Levels by Ethnicity

Analyze the Data

Suppose that one person from the study is randomly selected.

Exercise:

Problem: Find the probability that person smoked 11-20 cigarettes per day.

Solution:

$\frac{35,065}{100,450}$

Exercise:

Problem: Find the probability that person was Latino.

Solution:

$\frac{19,969}{100,450}$

Discussion Questions

Exercise:

Problem:

In words, explain what it means to pick one person from the study and that person is “Japanese American **AND** smokes 21-30 cigarettes per day.” Also, find the probability.

Solution:

$\frac{4,715}{100,450}$

Exercise:

Problem:

In words, explain what it means to pick one person from the study and that person is “Japanese American **OR** smokes 21-30 cigarettes per day.” Also, find the probability.

Solution:

$\frac{36,636}{100,450}$

Exercise:

Problem:

In words, explain what it means to pick one person from the study and that person is “Japanese American **GIVEN** that person smokes 21-30 cigarettes per day.” Also, find the probability.

Solution:

$\frac{4715}{15,273}$

Exercise:

Problem: Prove that smoking level/day and ethnicity are dependent events.

Practice 2: Calculating Probabilities
This module allows students to practice using what they've learned about Probability. Students will apply their understanding of basic probability terms, calculate probabilities based on the data provided, and determine whether events are independent or mutually exclusive.

Student Learning Outcomes

- Students will define basic probability terms.
- Students will calculate probabilities.
- Students will determine whether two events are mutually exclusive or whether two events are independent.

Note: Use probability rules to solve the problems below. Show your work.

Given

48% of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. (*Source: http://field.com/fieldpollonline/subscribers/Rls2393.pdf*).
37.6% of all Californians are Latino (*Source: U.S. Census Bureau*).

In this problem, let:

- $\text{C} = \text{Californians (registered voters) preferring life in prison without parole over the death penalty f}$
- $\text{L} = \text{Latino Californians}$

Suppose that one Californian is randomly selected.

Analyze the Data

Exercise:

Problem: $P(C)$ =

Solution:

0.48

Exercise:

Problem: $P(L)$ =

Solution:

0.376

Exercise:

Problem: $P(C|L)$ =

Solution:

0.55

Exercise:

Problem: In words, what is " $C|L$"?

Exercise:

Problem: $P(L \text{ AND } C)$ =

Solution:

0.2068

Exercise:

Problem: In words, what is “L and C”?

Exercise:

Problem: Are L and C independent events? Show why or why not.

Solution:

No

Exercise:

Problem: $P(L \text{ OR } C)$ =

Solution:

0.6492

Exercise:

Problem: In words, what is “L or C”?

Exercise:

Problem: Are L and C mutually exclusive events? Show why or why not.

Solution:

No

Homework

Probability: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean and provides a number of homework exercises related to Probability with contributions from Roberta Bloom.

Exercise:

Problem:

Suppose that you have 8 cards. 5 are green and 3 are yellow. The 5 green cards are numbered 1, 2, 3, 4, and 5. The 3 yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

- **a**List the sample space.
- **b**$\mathrm{P(G)} =$
- **c**$\mathrm{P(G|E)} =$
- **d**$\mathrm{P(G\ AND\ E)} =$
- **e**$\mathrm{P(G\ OR\ E)} =$
- **f**Are G and E mutually exclusive? Justify your answer numerically.

Solution:

- **a**$\{\mathrm{G1, G2, G3, G4, G5, Y1, Y2, Y3}\}$
- **b**$\frac{5}{8}$
- **c**$\frac{2}{3}$
- **d** $\frac{2}{8}$
- **e** $\frac{6}{8}$
- **f**No

Exercise:

Problem:

Refer to the previous problem. Suppose that this time you randomly draw two cards, one at a time, and **with replacement**.

- $G_1 =$ first card is green
- $G_2 =$ second card is green

- **a**Draw a tree diagram of the situation.
- **b** $P(G_1 \text{ AND } G_2) =$
- **c** $P(\text{at least one green}) =$
- **d** $P(G_2 \mid G_1) =$
- **e**Are G_2 and G_1 independent events? Explain why or why not.

Exercise:

Problem:

Refer to the previous problems. Suppose that this time you randomly draw two cards, one at a time, and **without replacement**.

- G_1= first card is green
- G_2= second card is green

- **a**Draw a tree diagram of the situation.
- **b**$\mathrm{P}(G_1 \text{ AND } G_2) =$
- **c**$\mathrm{P}(\text{at least one green}) =$
- **d**$\mathrm{P}(G_2|G_1) =$
- **e**Are G_2 and G_1 independent events? Explain why or why not.

Solution:

- **b** $(\frac{5}{8})(\frac{4}{7})$
- **c** $(\frac{5}{8})(\frac{3}{7}) + (\frac{3}{8})(\frac{5}{7}) + (\frac{5}{8})(\frac{4}{7})$
- **d** $\frac{4}{7}$
- **e**No

Exercise:

Problem: Roll two fair dice. Each die has 6 faces.

- **a**List the sample space.
- **b**Let A be the event that either a 3 or 4 is rolled first, followed by an even number. Find $\mathrm{P(A)}$.
- **c**Let B be the event that the sum of the two rolls is at most 7. Find $\mathrm{P(B)}$.
- **d**In words, explain what “$\mathrm{P(A|B)}$” represents. Find $\mathrm{P(A|B)}$.
- **e**Are A and B mutually exclusive events? Explain your answer in 1 - 3 complete sentences, including numerical justification.
- **f**Are A and B independent events? Explain your answer in 1 - 3 complete sentences, including numerical justification.

Exercise:

Problem:

A special deck of cards has 10 cards. Four are green, three are blue, and three are red. When a card is picked, the color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

- **a**List the sample space.

- **b**Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $\mathrm{P(A)}$.
- **c**Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in 1 - 3 complete sentences, including numerical justification.
- **d**Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in 1 - 3 complete sentences, including numerical justification.

Solution:

- **a** $\{\mathrm{GH,GT,BH,BT,RH,RT}\}$
- **b** $\frac{3}{20}$
- **c**Yes
- **d**No

Exercise:

Problem: An experiment consists of first rolling a die and then tossing a coin:

- **a**List the sample space.
- **b**Let A be the event that either a 3 or 4 is rolled first, followed by landing a head on the coin toss. Find $\mathrm{P(A)}$.
- **c**Let B be the event that a number less than 2 is rolled, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in 1 - 3 complete sentences, including numerical justification.

Exercise:

Problem:

An experiment consists of tossing a nickel, a dime and a quarter. Of interest is the side the coin lands on.

- **a**List the sample space.
- **b**Let A be the event that there are at least two tails. Find $\mathrm{P(A)}$.
- **c**Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in 1 - 3 complete sentences, including justification.

Solution:

- **a** $\{\mathrm{(HHH),(HHT),(HTH),(HTT),(THH),(THT),(TTH),(TTT)}\}$
- **b** $\frac{4}{8}$
- **c**Yes

Exercise:

Problem: Consider the following scenario:

- Let $\mathrm{P(C) = 0.4}$
- Let $\mathrm{P(D) = 0.5}$
- Let $\mathrm{P(C|D) = 0.6}$

- **a**Find $\mathrm{P(C\ AND\ D)}$.
- **b**Are C and D mutually exclusive? Why or why not?
- **c**Are C and D independent events? Why or why not?
- **d**Find $\mathrm{P(C\ OR\ D)}$.
- **e**Find $\mathrm{P(D|C)}$.

Exercise:

Problem: E and F mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E \mid F)$.

Solution:

0

Exercise:

Problem: J and K are independent events. $\mathrm{P(J \mid K) = 0.3}$. Find $P(J)$.

Exercise:

Problem: U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

- **a**$\mathrm{P(U\ AND\ V)}$ =
- **b** $\mathrm{P(U \mid V)}$ =
- **c**$\mathrm{P(U\ OR\ V)}$ =

Solution:

- **a**0
- **b** 0
- **c**0.63

Exercise:

Problem:

Q and R are independent events. $\mathrm{P(Q) = 0.4}$; $\mathrm{P(Q\ AND\ R) = 0.1}$. Find $\mathrm{P(R)}$.

Exercise:

Problem: Y and Z are independent events.

- **a** Rewrite the basic Addition Rule $\mathrm{P(Y\ OR\ Z) = P(Y) + P(Z) - P(Y\ AND\ Z)}$ using the information that Y and Z are independent events.
- **b** Use the rewritten rule to find $\mathrm{P(Z)}$ if $\mathrm{P(Y\ OR\ Z) = 0.71}$ and $\mathrm{P(Y) = 0.42}$.

Solution:

- **b**0.5

Exercise:

Problem: G and H are mutually exclusive events. $P(G) = 0.5$; $P(H) = 0.3$

- **a**Explain why the following statement MUST be false: $P(H \mid G) = 0.4$.
- **b**Find: $\mathrm{P(H\ OR\ G)}$.
- **c**Are G and H independent or dependent events? Explain in a complete sentence.

Exercise:

Problem:

The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3059 documented cases of AIDS in the county. They were grouped into the following categories (*Source: Santa Clara County Public H.D.*):

	Homosexual/Bisexual	**IV Drug User***	**Heterosexual Contact**	**Other**	**Totals**
Female	0	70	136	49	____
Male	2146	463	60	135	____
Totals	____	____	____	____	____

* includes homosexual/bisexual IV drug users

Suppose one of the persons with AIDS in Santa Clara County is randomly selected. Compute the following:

- **a** P(person is female) =
- **b**P(person has a risk factor Heterosexual Contact) =
- **c**P(person is female OR has a risk factor of IV Drug User) =
- **d**P(person is female AND has a risk factor of Homosexual/Bisexual) =
- **e**P(person is male AND has a risk factor of IV Drug User) =
- **f**P(female GIVEN person got the disease from heterosexual contact) =
- **g**Construct a Venn Diagram. Make one group females and the other group heterosexual contact.

Solution:

The completed contingency table is as follows:

	Homosexual/Bisexual	**IV Drug User***	**Heterosexual Contact**	**Other**	**Totals**
Female	0	70	136	49	**255**
Male	2146	463	60	135	**2804**
Totals	**2146**	**533**	**196**	**184**	**3059**

* includes homosexual/bisexual IV drug users

- **a** $\frac{255}{3059}$
- **b** $\frac{196}{3059}$
- **c** $\frac{718}{3059}$
- **d**0
- **e** $\frac{463}{3059}$
- **f** $\frac{136}{196}$

Exercise:

Problem:

Solve these questions using probability rules. Do NOT use the contingency table above. 3059 cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

- **a**$\mathrm{P(person\ is\ female)} =$
- **b**$\mathrm{P(person\ obtained\ the\ disease\ through\ heterosexual\ contact)} =$
- **c**$\mathrm{P(female\ GIVEN\ person\ got\ the\ disease\ from\ heterosexual\ contact)} =$
- **d**Construct a Venn Diagram. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Exercise:

Problem:

The following table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	**Brown**	**Blond**	**Black**	**Red**	**Totals**
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

- **a**Complete the table above.
- **b**What is the probability that a randomly selected child will have wavy hair?
- **c**What is the probability that a randomly selected child will have either brown or blond hair?
- **d**What is the probability that a randomly selected child will have wavy brown hair?
- **e**What is the probability that a randomly selected child will have red hair, given that he has straight hair?
- **f**If B is the event of a child having brown hair, find the probability of the complement of B.
- **g**In words, what does the complement of B represent?

Solution:

- **b**$\frac{43}{215}$
- **c** $\frac{120}{215}$
- **d**$\frac{20}{215}$
- **e**$\frac{12}{172}$
- **f** $\frac{115}{215}$

Exercise:

Problem:

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into the following table.

Shirt#	**≤ 210**	**211-250**	**251-290**	**290≤**
1-33	21	5	0	0
34-66	6	18	7	4
66-99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

- **a**Find the probability that his shirt number is from 1 to 33.
- **b**Find the probability that he weighs at most 210 pounds.
- **c**Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
- **d**Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
- **e**Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.
- **f**If having a shirt number from 1 to 33 and weighing at most 210 pounds were independent events, then what should be true about $\mathrm{P}(\text{Shirt}\#\text{ 1-33} \mid \leq 210\text{ pounds})$?

Exercise:

Problem:

Approximately 281,000,000 people over age 5 live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish. (*Source: http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf*)

Let: E = speak English at home; E’ = speak another language at home; S = speak Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	**Answers**
a. P(E') =	i. 0.8043
b. P(E) =	ii. 0.623
c. P(S and E') =	iii. 0.1957
d. P(S|E') =	iv. 0.1219

Solution:

- **a**iii
- **b**i
- **c**iv
- **d**ii

Exercise:

Problem:

The probability that a male develops some form of cancer in his lifetime is 0.4567 (Source: American Cancer Society). The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51 (Source: USA Today). Some of the questions below do not have enough information for you to answer them. Write “not enough information” for those answers.

Let: C = a man develops cancer in his lifetime; P = man has at least one false positive

- **a**Construct a tree diagram of the situation.
- **b**$\mathrm{P(C)}$ =
- **c**$P(\mathrm{P|C})$ =
- **d**$P(\mathrm{P|C'})$ =
- **e**If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Exercise:

Problem:

In 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let $\mathrm{G = won\ Green\ Card}$.

- **a**What was Renate’s chance of winning a Green Card? Write your answer as a probability statement.

- **b**In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Let $\mathrm{F} = \text{was a finalist}$. Write your answer as a conditional probability statement.
- **c**Are G and F independent or dependent events? Justify your answer numerically and also explain why.
- **d**Are G and F mutually exclusive events? Justify your answer numerically and also explain why.

Note:P.S. Amazingly, on 2/1/95, Renate learned that she would receive her Green Card -- true story!

Solution:

- **a** $P(G) = 0.008$
- **b** 0.5
- **c**dependent
- **d**No

Exercise:

Problem:

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned. (*Source: Wall Street Journal*)

Let: R = money returned; E = economics classes; O = other classes

- **a**Write a probability statement for the overall percent of money returned.
- **b**Write a probability statement for the percent of money returned out of the economics classes.
- **c**Write a probability statement for the percent of money returned out of the other classes.
- **d**Is money being returned independent of the class? Justify your answer numerically and explain it.
- **e**Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

Exercise:

Problem:

The chart below gives the number of suicides estimated in the U.S. for a recent year by age, race (black and white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population. (*Source: The National Center for Health Statistics, U.S. Dept. of Health and Human Services*)

Race and Sex	**1 - 14**	**15 - 24**	**25 - 64**	**over 64**	**TOTALS**
white, male	210	3360	13,610		22,050
white, female	80	580	3380		4930
black, male	10	460	1060		1670
black, female	0	40	270		330
all others					
TOTALS	310	4650	18,780		29,760

Note: Do not include "all others" for parts (f), (g), and (i).

- **a**Fill in the column for the suicides for individuals over age 64.
- **b**Fill in the row for all other races.
- **c**Find the probability that a randomly selected individual was a white male.
- **d**Find the probability that a randomly selected individual was a black female.
- **e**Find the probability that a randomly selected individual was black
- **f**Find the probability that a randomly selected individual was male.
- **g**Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.
- **h**Comparing “Race and Sex” to “Age,” which two groups are mutually exclusive? How do you know?
- **i**Are being male and committing suicide over age 64 independent events? How do you know?

Solution:

- **c** $\frac{22050}{29760}$
- **d** $\frac{330}{29760}$
- **e** $\frac{2000}{29760}$
- **f** $\frac{23720}{29760}$
- **g** $\frac{5010}{6020}$
- **h**Black females and ages 1-14
- **i**No

The next two questions refer to the following: The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20 - 64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20 - 64; 13.53% are age 65 or over. (Source: Federal Highway Administration, U.S. Dept. of Transportation)

Exercise:

Problem: Complete the following:

- **a**Construct a table or a tree diagram of the situation.
- **b**$\text{P(driver is female)}$ =
- **c**$\text{P(driver is age 65 or over} \mid \text{driver is female)}$ =
- **d**$\text{P(driver is age 65 or over AND female)}$ =
- **e**In words, explain the difference between the probabilities in part (c) and part (d).
- **f**$\text{P(driver is age 65 or over)}$ =
- **g**Are being age 65 or over and being female mutually exclusive events? How do you know

Exercise:

Problem: Suppose that 10,000 U.S. licensed drivers are randomly selected.

- **a**How many would you expect to be male?
- **b**Using the table or tree diagram from the previous exercise, construct a contingency table of gender versus age group.
- **c**Using the contingency table, find the probability that out of the age 20 - 64 group, a randomly selected driver is female.

Solution:

- **a**5140
- **c**0.49

Exercise:

Problem:

Approximately 86.5% of Americans commute to work by car, truck or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation. (*Source: Bureau of the Census, U.S. Dept. of Commerce. Disregard rounding approximations.*)

- **a**Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
- **b**Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
- **c**Suppose that 1000 workers are randomly selected. How many would you expect to travel alone to work?
- **d**Suppose that 1000 workers are randomly selected. How many would you expect to drive in a carpool?

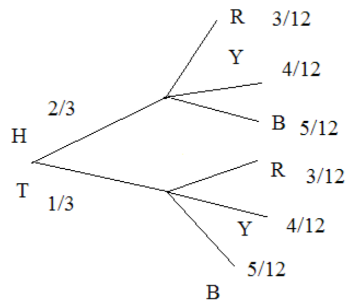
Exercise:

Problem: Explain what is wrong with the following statements. Use complete sentences.

- **a**If there's a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there's a 130% chance of rain over the weekend.
- **b**The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Try these multiple choice questions.

The next two questions refer to the following probability tree diagram which shows tossing an unfair coin **FOLLOWED BY** drawing one bead from a cup containing 3 red (R), 4 yellow (Y) and 5 blue (B) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where $\mathrm{H} = \text{"heads"}$ and $\mathrm{T} = \text{"tails"}$.



Exercise:

Problem: Find $\mathrm{P(tossing\ a\ Head\ on\ the\ coin\ AND\ a\ Red\ bead)}$

- **A** $\frac{2}{3}$
- **B** $\frac{5}{15}$
- **C** $\frac{6}{36}$
- **D** $\frac{5}{36}$

Solution:

C

Exercise:

Problem: Find $\mathrm{P(Blue\ bead)}$.

- **A** $\frac{15}{36}$
- **B** $\frac{10}{36}$
- **C** $\frac{10}{12}$
- **D** $\frac{6}{36}$

Solution:

A

The next three questions refer to the following table of data obtained from *www.baseball-almanac.com* showing hit information for 4 well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1517	506	136	714	2873
Jackie Robinson	1054	273	54	137	1518
Ty Cobb	3603	174	295	114	4189
Hank Aaron	2294	624	98	755	3771

NAME	Single	Double	Triple	Home Run	TOTAL HITS
TOTAL	8471	1577	583	1720	12351

Exercise:

Problem: Find $\text{P(hit was made by Babe Ruth)}$.

- **A** $\frac{1518}{2873}$
- **B** $\frac{2873}{12351}$
- **C** $\frac{583}{12351}$
- **D** $\frac{4189}{12351}$

Solution:

B

Exercise:

Problem: Find $\text{P(hit was made by Ty Cobb | The hit was a Home Run)}$

- **A** $\frac{4189}{12351}$
- **B** $\frac{114}{1720}$
- **C** $\frac{1720}{4189}$
- **D** $\frac{114}{12351}$

Solution:

B

Exercise:

Problem:

Are the hit being made by Hank Aaron and the hit being a double independent events?

- **A** Yes, because $\text{P(hit by Hank Aaron | hit is a double)} = \text{P(hit by Hank Aaron)}$
- **B** No, because $\text{P(hit by Hank Aaron | hit is a double)} \neq \text{P(hit is a double)}$
- **C** No, because $\text{P(hit is by Hank Aaron | hit is a double)} \neq \text{P(hit by Hank Aaron)}$
- **D** Yes, because $\text{P(hit is by Hank Aaron | hit is a double)} = \text{P(hit is a double)}$

Solution:

C

Exercise:

Problem: Given events G and H: P(G) = 0.43 ; P(H) = 0.26 ; P(H and G) = 0.14

- **A** Find P(H or G)
- **B**Find the probability of the complement of event (H and G)
- **C**Find the probability of the complement of event (H or G)

Solution:

- **A** P(H or G) = P(H) + P(G) − P(H and G) = 0.26 + 0.43 − 0.14 = 0.55
- **B** P(NOT (H and G)) = 1 − P(H and G) = 1 − 0.14 = 0.86
- **C** P(NOT (H or G)) = 1 − P(H or G) = 1 − 0.55 = 0.45

Exercise:

Problem: Given events J and K: P(J) = 0.18 ; P(K) = 0.37 ; P(J or K) = 0.45

- **A** Find P(J and K)
- **B** Find the probability of the complement of event (J and K)
- **C** Find the probability of the complement of event (J or K)

Solution:

- **A**P(J or K) = P(J) + P(K) − P(J and K); 0.45 = 0.18 + 0.37 − P(J and K) ; solve to find P(J and K) = 0.10
- **B** P(NOT (J and K)) = 1 − P(J and K) = 1 − 0.10 = 0.90
- **C** P(NOT (J or K)) = 1 − P(J or K) = 1 − 0.45 = 0.55

Exercise:

Problem:

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, http://www.unitedbloodservices.org/humanbloodtypes.html, a person with type O blood and a negative Rh factor (Rh−) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh− factor; 52% of people have type O or Rh− factor.

- **A** Find the probability that a person has both type O blood and the Rh− factor
- **B** Find the probability that a person does NOT have both type O blood and the Rh− factor.

Solution:

- **A** P(Type O or Rh−) = P(Type O) + P(Rh−) − P(Type O and Rh−)
  0.52 = 0.43 + 0.15 − P(Type O and Rh−); solve to find P(Type O and Rh−) = 0.06
  6% of people have type O Rh− blood
- **B** P(NOT (Type O and Rh−)) = 1 − P(Type O and Rh−) = 1 − 0.06 = 0.94
  94% of people do not have type O Rh− blood

Exercise:

Problem:

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

- **A** Find the probability that a course has a final exam or a research project.
- **B** Find the probability that a course has NEITHER of these two requirements.

Solution:

- **A** P(R or F) = P(R) + P(F) − P(R and F) = 0.72 + 0.46 − 0.32 = 0.86
- **B** P(Neither R nor F) = 1 − P(R or F) = 1 − 0.86 = 0.14

Exercise:

Problem:

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

- **A** Find the probability that a cookie contains chocolate or nuts (he can't eat it).
- **B** Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

Solution:

- Let C be the event that the cookie contains chocolate. Let N be the event that the cookie contains nuts.
- **A** P(C or N) = P(C) + P(N) − P(C and N) = 0.36 + 0.12 − 0.08 = 0.40
- **B** P(neither chocolate nor nuts) = 1 − P(C or N) = 1 − 0.40 = 0.60

Exercise:

Problem:

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student

- **A** Find P(D and E)
- **B** Find P(E | D)
- **C** Find P(D or E)
- **D** Using an appropriate test, show whether D and E are independent.
- **E** Using an appropriate test, show whether D and E are mutually exclusive.

Solution:

- **A** P(D and E) = P(D|E)P(E) = (0.20)(0.40) = 0.08
- **B** P(E|D) = P(D and E) / P(D) = 0.08/0.10 = 0.80
- **C** P(D or E) = P(D) + P(E) − P(D and E) = 0.10 + 0.40 − 0.08 = 0.42
- **D** Not Independent: P(D|E) = 0.20 which does not equal P(D) = .10
- **E** Not Mutually Exclusive: P(D and E) = 0.08 ; if they were mutually exclusive then we would need to have P(D and E) = 0, which is not true here.

Exercise:

Problem:

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian 1 Euro coin was a fair coin. They spun the coin rather than tossing it, and it was found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). Therefore, they claim that this is not a fair coin.

- **A** Based on the data above, find P(H) and P(T).
- **B** Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
- **C** Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
- **D** Use the tree to find the probability of obtaining at least one head.

Solution:

- **A** P(H) = 140/250; P(T) = 110/250
- **C**308/625
- **D** 504/625

Exercise:

Problem:

A box of cookies contains 3 chocolate and 7 butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it also. (How many cookies did he take?)

- **A** Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.

- **B** Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
- **C** For each complete path through the tree, write the event it represents and find the probabilities.
- **D** Let S be the event that both cookies selected were the same flavor. Find P(S).
- **E** Let T be the event that both cookies selected were different flavors. Find P(T) by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
- **F** Let U be the event that the second cookie selected is a butter cookie. Find P(U).

**Exercises 33 - 40 contributed by Roberta Bloom

Review

This module provides a number of homework/review exercises related to Probability.

The first six exercises refer to the following study: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks. Answer the following:

Exercise:

Problem: The “average increase” for all NASDAQ stocks is the:

- **A**Population
- **B**Statistic
- **C**Parameter
- **D**Sample
- **E**Variable

Solution:

- **C** Parameter

Exercise:

Problem: All of the NASDAQ stocks are the:

- **A**Population
- **B**Statistic
- **C**Parameter
- **D**Sample
- **E**Variable

Solution:

- **A** Population

Exercise:

Problem: 9% is the:

- **A**Population
- **B**Statistic
- **C**Parameter
- **D**Sample
- **E**Variable

Solution:

- **B** Statistic

Exercise:

Problem: The 100 NASDAQ stocks in the survey are the:

- **A**Population
- **B**Statistic
- **C**Parameter
- **D**Sample
- **E**Variable

Solution:

- **D** Sample

Exercise:

Problem: The percent increase for one stock in the survey is the:

- **A**Population
- **B**Statistic

- **C**Parameter
- **D**Sample
- **E**Variable

Solution:

- **E** Variable

Exercise:

Problem:

Would the data collected be qualitative, quantitative – discrete, or quantitative – continuous?

Solution:

quantitative - continuous

The next two questions refer to the following study: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Weight Gain	Frequency
-2	3
-1	5
0	2

Weight Gain	**Frequency**
1	4
4	13
6	2
11	1

Exercise:

Problem: Calculate the following values:

- **a**The average weight gain for the two weeks
- **b**The standard deviation
- **c**The first, second, and third quartiles

Solution:

- **a** 2.27
- **b** 3.04
- **c** -1, 4, 4

Exercise:

Problem: Construct a histogram and a boxplot of the data.

Introduction
This module serves as the introduction to Discrete Random Variables in the Elementary Statistics textbook/collection.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately (optional).
- Recognize the geometric probability distribution and apply it appropriately (optional).
- Recognize the hypergeometric probability distribution and apply it appropriately (optional).
- Classify discrete word problems by their distributions.

Introduction

A student takes a 10 question true-false quiz. Because the student had such a busy schedule, he or she could not study and randomly guesses at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

In this chapter, you will study probability problems involving discrete random distributions. You will also study long-term averages associated with them.

Random Variable Notation

Upper case letters like X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If X **is a random variable, then** X **is written in words.** and x **is given as a number.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is $\text{TTT THH HTH HHT HTT THT TTH HHH}$. Then, x = 0, 1, 2, 3. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Optional Collaborative Classroom Activity

Toss a coin 10 times and record the number of heads. After all members of the class have completed the experiment (tossed a coin 10 times and counted the number of heads), fill in the chart using a heading like the one below. Let X = the number of heads in 10 tosses of the coin.

x	**Frequency of x**	**Relative Frequency of x**

x	**Frequency of x**	**Relative Frequency of x**

- Which value(s) of x occurred most frequently?
- If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
- What does the relative frequency column sum to?

Glossary

Random Variable (RV)

see **Variable**

Variable (Random Variable)

A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

Probability Distribution Function (PDF) for a Discrete Random Variable
This module introduces the Probability Distribution Function (PDF) and its characteristics.

A discrete **probability distribution function** has two characteristics:

- Each probability is between 0 and 1, inclusive.
- The sum of the probabilities is 1.

Example:
A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times a newborn wakes its mother after midnight. For this example, x = 0, 1, 2, 3, 4, 5.
$\mathrm{P(x)}$ = probability that X takes on a value x.

x	$\mathrm{P(x)}$
0	$\mathrm{P(x=0)} = \frac{2}{50}$
1	$\mathrm{P(x=1)} = \frac{11}{50}$
2	$\mathrm{P(x=2)} = \frac{23}{50}$
3	$\mathrm{P(x=3)} = \frac{9}{50}$
4	$\mathrm{P(x=4)} = \frac{4}{50}$
5	

	$\mathrm{P(x{=}5)} = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because

1. Each $\mathrm{P(x)}$ is between 0 and 1, inclusive.
2. The sum of the probabilities is 1, that is,

Equation:

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

Example:
Suppose Nancy has classes **3 days** a week. She attends classes 3 days a week **80%** of the time, **2 days 15%** of the time, **1 day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Exercise:

Problem:

Let X = the number of days Nancy ______________________ .

Solution:

Let X = the number of days Nancy **attends class per week**.

Exercise:

Problem: X takes on what values?

Solution:

0, 1, 2, and 3

Exercise:

Problem:

Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in the previous example. The table should have two columns labeled x and $\mathrm{P}(\mathrm{x})$. What does the $\mathrm{P}(\mathrm{x})$ column sum to?

Solution:

x	$\mathrm{P}(\mathrm{x})$
0	0.01
1	0.04
2	0.15
3	0.80

Glossary

Probability Distribution Function (PDF)

A mathematical description of a discrete random variable (RV), given either in the form of an equation (formula) , or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Example:

A biased coin with probability 0.7 for a head (in one toss of the coin) is tossed 5 times. We are interested in the number of heads (the RV X = the number of heads). X is Binomial, so $X \sim B(5, 0.7)$ and $P(X = x) = \binom{5}{x}.7^{x}.3^{5-x}$or in the form of the table:

x	$P(X = x)$
0	0.0024
1	0.0284
2	0.1323
3	0.3087
4	0.3602
5	0.1681

Mean or Expected Value and Standard Deviation
This module explores the Law of Large Numbers, the phenomenon where an experiment performed many times will yield cumulative results closer and closer to the theoretical mean over time.

The **expected value** is often referred to as the **"long-term"average or mean** . This means that over the long term of doing an experiment over and over, you would **expect** this average.

The **mean** of a random variable X is μ. If we do an experiment many times (for instance, flip a fair coin, as Karl Pearson did, 24,000 times and let X = the number of heads) and record the value of X each time, the average is likely to get closer and closer to μ as we keep repeating the experiment. This is known as the **Law of Large Numbers**.

Note: To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

A Step-by-Step Example
A men's soccer team plays soccer 0, 1, or 2 days a week. The probability that they play 0 days is 0.2, the probability that they play 1 day is 0.5, and the probability that they play 2 days is 0.3. Find the long-term average, μ, or expected value of the days per week the men's soccer team plays soccer.

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table, adding a column $\mathrm{xP(x)}$. In this column, you will multiply each x value by its probability.

x	$\mathrm{P(x)}$	$x\mathrm{P(x)}$
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Expected Value TableThis table is called an expected value table. The table helps you calculate the expected value or long-term average.

Add the last column to find the long term average or expected value: $(0)(0.2)+(1)(0.5)+(2)(0.3)= 0 + 0.5 + 0.6 = 1.1.$

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu=1.1$

Example:
Find the expected value for the example about the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times a newborn wakes its mother after midnight.

x	$\mathrm{P(X)}$	$x\mathrm{P(X)}$
0	$\mathrm{P(x=0)} = \frac{2}{50}$	(0)$\left(\frac{2}{50}\right)$ = 0

x	P(X)	xP(X)
1	$\text{P}(\text{x=1}) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$
2	$\text{P}(\text{x=2}) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$
3	$\text{P}(\text{x=3}) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$
4	$\text{P}(\text{x=4}) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$
5	$\text{P}(\text{x=5}) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$

You expect a newborn to wake its mother after midnight 2.1 times, on the average.

Add the last column to find the expected value. μ = Expected Value = $\frac{105}{50} = 2.1$

Exercise:

Problem:

Go back and calculate the expected value for the number of days Nancy attends classes a week. Construct the third column to do so.

Solution:

2.74 days a week.

Example:

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from 0 to 9 with replacement. You pay $2 to play and could profit $100,000 if you match all 5 numbers in order (you get your $2 back plus

$100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and -2 dollars.

To win, you must get all 5 numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are 10 numbers. You may choose a number more than once. The probability of choosing all 5 numbers correctly and in order is:

Equation:

$$\frac{1}{10}*\frac{1}{10}*\frac{1}{10}*\frac{1}{10}*\frac{1}{10}*= 1*10^{-5} = 0.00001$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

Equation:

$$1 - 0.00001 = 0.99999$$

The expected value table is as follows.

	x	$\mathrm{P(x)}$	$x\mathrm{P(x)}$
Loss	-2	0.99999	(-2)(0.99999)=-1.99998
Profit	100,000	0.00001	(100000)(0.00001)=1

Add the last column. -1.99998 + 1 = -0.99998

Since -0.99998 is about -1, you would, on the average, expect to lose approximately one dollar for each game you play. However, each time you play, you either lose \$2 or profit \$100,000. The \$1 is the average or expected LOSS per game after playing this game over and over.

Example:
Suppose you play a game with a biased coin. You play each game by tossing the coin once. $\mathrm{P}(\text{heads}) = \frac{2}{3}$ and $\mathrm{P}(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead?

Exercise:

Problem: Define a random variable X.

Solution:

X = amount of profit

Exercise:

Problem: Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

Solution:

	x	$\mathrm{P(x)}$	$x\mathrm{P(x)}$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	**-6**	$\frac{2}{3}$	$\frac{-12}{3}$

Exercise:

Problem: What is the expected value, μ? Do you come out ahead?

Solution:

Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root . To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x - \mu)^2 \cdot P\left(x\right)$ and take the square root.

x	$\mathrm{P(x)}$	$x\mathrm{P(x)}$	$(\mathrm{x}\ \text{-}\mu)^2\mathrm{P(x)}$
0	0.2	(0)(0.2) = 0	$(0-1.1)^2\left(.2\right)=0.242$
1	0.5	(1)(0.5) = 0.5	$(1-1.1)^2\left(.5\right)=0.005$
2	0.3	(2)(0.3) = 0.6	$(2-1.1)^2\left(.3\right)=0.243$

Add the last column in the table. $0.242+0.005+0.243=0.490$. The standard deviation is the square root of 0.49. $\sigma=\sqrt{0.49}=0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas that calculate μ and σ.

Glossary

Expected Value

Expected arithmetic average when an experiment is repeated many times. (Also called the mean). Notations: $E(x)$, μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $E(x)=\mu=\sum\mathrm{xP}(x)$.

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x=\frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu=\frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Common Discrete Probability Distribution Functions

This module serves as a lead-in for several types of common discrete probability distribution functions, including binomial, geometric, hypergeometric, and Poisson.

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

The Binomial Random Variable

When you flip a coin, there are two possible outcomes: heads and tails. Each outcome has a fixed probability, the same from trial to trial. In the case of coins, heads and tails each have the same probability of 1/2. More generally, there are situations in which the coin is biased, so that heads and tails have different probabilities. In the present section, we consider probability distributions for which there are just two possible outcomes with fixed probability summing to one. These distributions are called are called **binomial distributions**.

A Simple Example

The four possible outcomes that could occur if you flipped a coin twice are listed in [link]. Note that the four outcomes are equally likely: each has probability $1/4$. To see this, note that the tosses of the coin are independent (neither affects the other). Hence, the probability of a head on Flip 1 and a head on Flip 2 is the product of $\mathrm{P[H]}$ and $\mathrm{P[H]}$, which is $1/2 \times 1/2 = 1/4$. The same calculation applies to the probability of a head on Flip one and a tail on Flip 2. Each is $1/2 \times 1/2 = 1/4$.

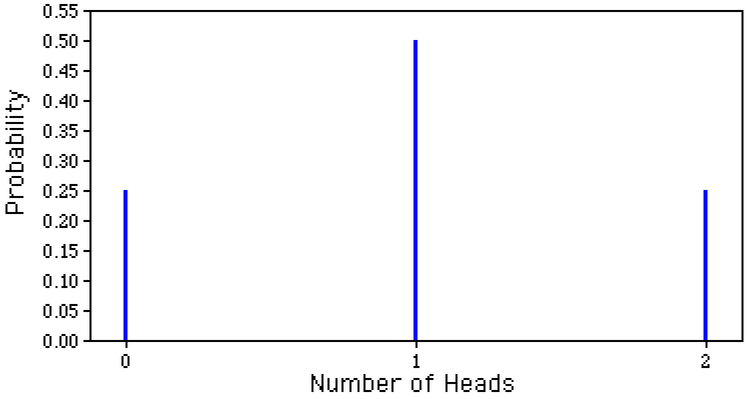
Outcome	**First Flip**	**Second Flip**
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails

Four Possible Outcomes

The four possible outcomes can be classifid in terms of the number of heads that come up. The number could be two (Outcome 1), one (Outcomes 2 and 3) or 0 (Outcome 4). The probabilities of these possibilities are shown in [link] and in [link]. Since two of the outcomes represent the case in which just one head appears in the two tosses, the probability of this event is equal to $1/4+1/4=1/2$. [link] summarizes the situation.

Number of Heads	Probability
0	1/4
1	1/2
2	1/4

Probabilities of Getting 0,1, or 2 heads.



Probabilities of 0, 1, and 2 heads.

[link] is a discrete probability distribution: It shows the probability for each of the values on the X-axis. Defining a head as a "success," [link] shows the probability of 0, 1, and 2 successes for two trials (flips) for an event that has a probability of 0.5 of being a success on each trial. This makes [link] an example of a **binomial distribution**.

The Formula for Binomial Probabilities

The binomial distribution consists of the probabilities of each of the possible numbers of successes on n trials for independent events that each have a probability of p of occurring. For the coin flip example, $n = 2$ and $\text{p}=(0.5)$. The formula for the binomial distribution is shown below:

$$\text{P}[\text{x}] = \frac{n!}{x!\,(n-x)!}p^{x}(1-p)^{n-x}$$

where $\text{P}[\text{x}]$ is the probability of x successes out of n trials, n is the number of trials, and p is the probability of success on a given trial. Applying this to the coin flip example,

$$\text{P}[0] = \frac{2!}{0!\,(2-0)!}0.5^{0}(1-0.5)^{2-0} = \frac{2}{2}1 \times .25 = 0.25$$

$$\text{P}[1] = \frac{2!}{1!\,(2-1)!}0.5^{1}(1-0.5)^{2-1} = \frac{2}{1}.5 \times .5 = 0.50$$

$$\text{P}[2] = \frac{2!}{2!\,(2-2)!}0.5^{2}(1-0.5)^{2-2} = \frac{2}{2}.25 \times 1 = 0.25$$

If you flip a coin twice, what is the probability of getting one or more heads? Since the probability of getting exactly one head is 0.50 and the probability of getting exactly two heads is 0.25, the probability of getting one or more heads is $0.50 + 0.25 = 0.75$.

Now suppose that the coin is biased; let's say the probability of heads is only 0.4. What is the probability of getting heads at least once in two tosses? We could substitute p=0.4 with x=1 and with x=2 into our general formula above; adding the results would obtain the answer 0.64.

Cumulative Probabilities

We toss a coin 12 times. What is the probability that we get from 0 to 3 heads? The answer is found by computing the probability of exactly 0 heads, exactly 1 head, exactly 2 heads, and exactly 3 heads. The probability of getting from 0 to 3 heads is then the sum of these probabilities. The probabilities are: 0.0002, 0.0029, 0.0161, and 0.0537. The sum of the probabilities is 0.073. The calculation of cumulative binomial probabilities can be quite tedious. Therefore we have provided a binomial calculator to make it easy to calculate these probabilities.

Note: Click here for the binomial calculator.

Mean and Standard Deviation of Binomial Distributions

Consider a coin-tossing experiment in which you tossed a coin 12 times and recorded the number of heads. If you performed this experiment over and over again, what would the mean number of heads be? On average, you would expect half the coin tosses to come up heads. Therefore the mean number of heads would be 6. In general, the mean of a binomial distribution with parameters n (the number of trials) and p (the probability of success for each trial) is:

$$\mu = np$$

where μ is the mean of the binomial distribution. The variance of the binomial distribution is:

$$\sigma^2 = np\left(1 - p\right)$$

where σ^2 is the variance of the binomial distribution.

Let's return to the coin tossing experiment. The coin was tossed 12 times so $n = 12$. A coin has a probability of 0.5 of coming up heads. Therefore, $p = 0.5$. The mean and standard deviation can therefore be computed as follows:

$$\mu = np = 12 \times 0.5 = 6$$

$$\sigma^2 = np\left(1 - p\right) = 12 \times 0.5 \times \left(1.0 - 0.5\right) = 3.0$$

Naturally, the standard deviation $\left(\sigma\right)$ is the square root of the variance $\left(\sigma^2\right)$.

Binomial Calculator

Applet failed to run. No Java plug-in was found.

Binomial Experiments

This module describes the characteristics of a binomial experiment and the binomial probability distribution function.

The characteristics of a binomial experiment are:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only 2 possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true - false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true - false question with probability $p = 0.6$. Then, $q = 0.4$ .This means that for every true - false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the n independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution is $\mu = \mathrm{np}$ and $\sigma^2 = \mathrm{npq}$. The standard deviation, σ, is then $\sigma = \sqrt{\mathrm{npq}}$.

Any experiment that has characteristics 2 and 3 and where n = 1 is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example:
At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable is X = the number of students who withdraw from the randomly selected elementary physics class.

Example:
Suppose you play a game that you can only either win or lose. The probability that you win any game is 55% and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, what is the probability that you win 15 of the 20 games? Here, if you define X = the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Example:
A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than 10 heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, p = 0.5 and q = 0.5. The number of trials is n = 15. The probability question can be stated mathematically as $P(x > 10)$.

Example:
Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Exercise:

Problem:

This is a binomial problem because there is only a success or a __________, there are a definite number of trials, and the probability of a success is 0.70 for each trial.

Solution:

failure

Exercise:

Problem:

If we are interested in the number of students who do their homework, then how do we define X?

Solution:

X = the number of statistics students who do their homework on time

Exercise:

Problem: What values does x take on?

Solution:

0, 1, 2, …, 50

Exercise:

Problem: What is a "failure", in words?

Solution:

Failure is a student who does not do his or her homework on time.

The probability of a success is p = 0.70. The number of trial is n = 50.

Exercise:

Problem: If $p + q = 1$, then what is q?

Solution:

q = 0.30

Exercise:

Problem:

The words "at least" translate as what kind of inequality for the probability question $P(x\_\_\_\_40)$.

Solution:

greater than or equal to (≥)

The probability question is $P(x \geq 40)$.

Notation for the Binomial: B = Binomial Probability Distribution Function

X ~ $B(n, p)$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p. n = number of trials p = probability of a success on each trial

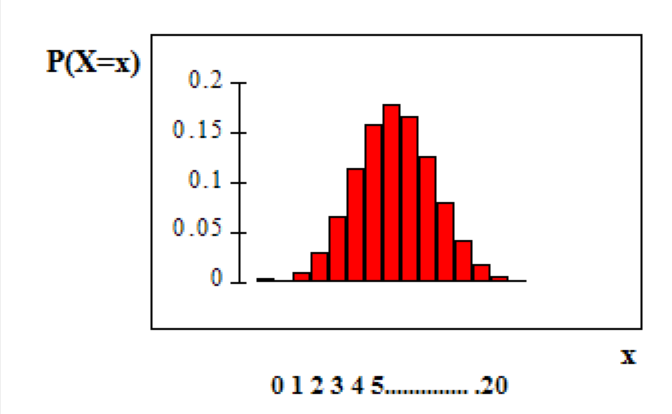
Example:

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult

workers do you expect to have a high school diploma but do not pursue any further education?
Let X = the number of workers who have a high school diploma but do not pursue any further education.
X takes on the values 0, 1, 2, ..., 20 where n = 20 and p = 0.41. q = 1 - 0.41 = 0.59. X ~ $B(20, 0.41)$
Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)
Using the TI-83+ or the TI-84 calculators, the calculations are as follows. Go into 2nd DISTR. The syntax for the instructions are
To calculate (x = value): binompdf(n, p, number) If "number" is left out, the result is the binomial probability table.
To calculate $P(x \leq \text{value})$: binomcdf(n, p, number) If "number" is left out, the result is the cumulative binomial probability table.
For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,.41,12). The result is $P(x \leq 12) = 0.9738$.

Note: If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find P(x>12), use 1 - binomcdf(20,.41,12).

The probability at most 12 workers have a high school diploma but do not pursue any further education is 0.9738
The graph of x ~ $B(20, 0.41)$ is:



The y-axis contains the probability of x, where X = the number of workers who have only a high school diploma.
The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = \text{np} = (20)(0.41) = 8.2$.
The formula for the variance is $\sigma^2 = \text{npq}$. The standard deviation is $\sigma = \sqrt{\text{npq}}$. $\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20$.

Example:
The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of 10 staff members and 6 students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? All names of the committee are put into a box and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw produces a student. The probability is $\frac{6}{15}$ when the first draw produces a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Glossary

Bernoulli Trials

An experiment with the following characteristics:

- There are only 2 possible outcomes called “success” and “failure” for each trial.
- The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: X~ $B(n,p)$. The mean is $\mu = \mathrm{np}$ and the standard deviation is $\sigma = \sqrt{\mathrm{npq}}$. The probability of exactly x successes in n trials is $P\left(X = x\right) = \binom{n}{x}p^{x}q^{n-x}$.

Summary of the Binomial Formulas

This module provides a review of the binomial, geometric, hypergeometric, and Poisson probability distribution functions and their properties.

Formula

Binomial

X~$B(n, p)$

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x =$ 0,1, 2, 3, ...,n

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1 \qquad q = 1 - p$

The mean is $\mu = \mathrm{np}$. The standard deviation is $\sigma = \sqrt{\mathrm{npq}}$.

Practice 1: Discrete Distribution
This module provides students an opportunity to practice applying concepts related to discrete distributions. This practice exercise asks students to calculate several values based on the data provided.

Student Learning Outcomes

- The student will analyze the properties of a discrete distribution.

Given:

A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let $\mathrm{P}(\mathrm{x})$ = the probability that a student will study ballet x years.

Organize the Data

Complete the table below using the data provided.

x	**P(x)**	**x*P(x)**
1	0.10	
2	0.05	
3	0.10	

x	P(x)	x*P(x)
4		
5	0.30	
6	0.20	
7	0.10	

Exercise:

Problem: In words, define the Random Variable X.

Exercise:

Problem: $\mathrm{P}(\mathrm{x} = 4) =$

Exercise:

Problem: $P(x < 4) =$

Exercise:

Problem:

On average, how many years would you expect a child to study ballet with this teacher?

Discussion Question

Exercise:

Problem: What does the column "$\mathrm{P}(\mathrm{x})$" sum to and why?

Exercise:

Problem: What does the column "x*$\mathrm{P}(\mathrm{x})$" sum to and why?

Practice 2: Binomial Distribution
This module provides a practice of Binomial Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will construct the Binomial Distribution.

Given

The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. (*Source: http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf)*.)

Suppose that you randomly pick 8 first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status

Interpret the Data

Exercise:

Problem: In words, define the random Variable X.

Solution:

X= the number that reply “yes”

Exercise:

Problem: X~___________

Solution:

$B(8, 0.713)$

Exercise:

Problem: What values does the random variable X take on?

Solution:

0,1,2,3,4,5,6,7,8

Exercise:

Problem: Construct the probability distribution function (PDF).

x	P(x)

Exercise:

Problem:

On average (u), how many would you expect to answer yes?

Solution:

5.7

Exercise:

Problem: What is the standard deviation (σ) ?

Solution:

1.28

Exercise:

Problem:

What is the probability that at most 5 of the freshmen reply “yes”?

Solution:

0.4151

Exercise:

Problem:

What is the probability that at least 2 of the freshmen reply “yes”?

Solution:

0.9990

Exercise:

Problem:

Construct a histogram or plot a line graph. Label the horizontal and vertical axes with words. Include numerical scaling.

Homework
Discrete Random Variables: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean Homework and provides a number of homework exercises related to Discrete Random Variables (binomial, geometric, hypergeometric and Poisson) with contributions from Roberta Bloom.

Exercise:

Problem: 1. Complete the PDF and answer the questions.

x	$P(X = x)$	$x \cdot P(X = x)$
0	0.3	
1	0.2	
2		
3	0.4	

- **a**Find the probability that $x = 2$.
- **b** Find the expected value.

Solution:

- **a**0.1
- **b**1.6

Exercise:

Problem:

Suppose that you are offered the following "deal." You roll a die. If you roll a 6, you win $10. If you roll a 4 or 5, you win $5. If you roll a 1, 2, or 3, you pay $6.

- **a**What are you ultimately interested in here (the value of the roll or the money you win)?
- **b**In words, define the Random Variable X.
- **c**List the values that X may take on.
- **d**Construct a PDF.
- **e**Over the long run of playing this game, what are your expected average winnings per game?
- **f**Based on numerical values, should you take the deal? Explain your decision in complete sentences.

Exercise:

Problem:

A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

- **a**Construct a PDF for each investment.
- **b**Find the expected value for each investment.
- **c**Which is the safest investment? Why do you think so?
- **d**Which is the riskiest investment? Why do you think so?
- **e**Which investment has the highest expected return, on average?

Solution:

- **b**$200,000;$600,000;$400,000
- **c**third investment
- **d**first investment
- **e**second investment

Exercise:

Problem:

A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 apiece. Suppose you purchase 4 tickets. The prize is 2 passes to a Broadway show, worth a total of $150.

- **a**What are you interested in here?
- **b**In words, define the Random Variable X.
- **c**List the values that X may take on.
- **d**Construct a PDF.
- **e**If this fund-raiser is repeated often and you always purchase 4 tickets, what would be your expected average winnings per raffle?

Exercise:

Problem:

Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children

x	$P(X = x)$	$x \cdot P(X = x)$

0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

- **a**Find the probability that a married adult has 3 children.
- **b**In words, what does the expected value in this example represent?
- **c** Find the expected value.
- **d** Is it more likely that a married adult will have 2 – 3 children or 4 – 6 children? How do you know?

Solution:

- **a**0.2
- **c**2.35
- **d**2-3 children

Exercise:

Problem:

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given below.

x	$P(X = x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

- **a**In words, define the Random Variable X.
- **b** What does it mean that the values 0, 1, and 2 are not included for x in the PDF?
- **c**On average, how many years do you expect it to take for an individual to earn a B.S.?

For each problem:

- **a**In words, define the Random Variable X.
- **b**List the values that X may take on.
- **c**Give the distribution of X. X~

Then, answer the questions specific to each individual problem.

Exercise:

Problem:

Six different colored dice are rolled. Of interest is the number of dice that show a “1.”

- **d**On average, how many dice would you expect to show a “1”?
- **e**Find the probability that all six dice show a “1.”

- **f**Is it more likely that 3 or that 4 dice will show a “1”? Use numbers to justify your answer numerically.

Solution:

- **a** X = the number of dice that show a 1
- **b**0,1,2,3,4,5,6
- **c** $X \sim B\left(6, \frac{1}{6}\right)$
- **d** 1
- **e** 0.00002
- **f** 3 dice

Exercise:

Problem:

More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses. *(Source: http://en.wikipedia.org/wiki/Distance_education)*

- **d**On average, how many schools would you expect to offer such courses?
- **e**Find the probability that at most 6 offer such courses.
- **f**Is it more likely that 0 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

Exercise:

Problem:

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

- **d**How many of the 12 students do we expect to attend the festivities?
- **e**Find the probability that at most 4 students will attend.
- **f**Find the probability that more than 2 students will attend.

Solution:

- **a** X = the number of students that will attend Tet.
- **b**0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
- **c** $X \sim \mathrm{B}(12{,}0.18)$
- **d**2.16
- **e**0.9511
- **f**0.3702

Exercise:

Problem:

Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

- **d**How many are expected to attend their graduation?
- **e**Find the probability that 17 or 18 attend.
- **f**Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Exercise:

Problem:

At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the numbers that do **not** use the foil as their main weapon.

- **d**How many are expected to **not** use the foil as their main weapon?
- **e**Find the probability that six do **not** use the foil as their main weapon.
- **f**Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

Solution:

- **a** X = the number of fencers that do **not** use foil as their main weapon
- **b**0, 1, 2, 3,... 25
- **c** X~$\text{B}(25,0.40)$
- **d**10
- **e**0.0442
- **f**Yes

Exercise:

Problem:

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.

- **d**How many seniors are expected to have participated in after-school sports all four years of high school?

- **e**Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
- **f**Based upon numerical values, is it more likely that 4 or that 5 of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Exercise:

Problem:

The chance of having an extra fortune in a fortune cookie is about 3%. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem. Use one distribution to solve the problem.

- **d**How many cookies do we expect to have an extra fortune?
- **e**Find the probability that none of the cookies have an extra fortune.
- **f**Find the probability that more than 3 have an extra fortune.
- **g**As n increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

Solution:

- **a** X = the number of fortune cookies that have an extra fortune
- **b**0, 1, 2, 3,... 144
- **c** X~$\text{B}(144, 0.03)$ or $\text{P}(4.32)$
- **d** 4.32
- **e** 0.0124 or 0.0133
- **f** 0.6300 or 0.6264

Exercise:

Problem:

There are two games played for Chinese New Year and Vietnamese New Year. They are almost identical. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his $1 bet, plus $3 profit.

Let X = number of matches and Y= profit per game.

- **d**List the values that Y may take on. Then, construct one PDF table that includes both X & Y and their probabilities.
- **e**Calculate the average expected matches over the long run of playing this game for the player.
- **f**Calculate the average expected earnings over the long run of playing this game for the player.
- **g**Determine who has the advantage, the player or the house.

Exercise:

Problem:

According to the South Carolina Department of Mental Health web site, for every 200 U.S. women, the average number who suffer from anorexia is one (*http://www.state.sc.us/dmh/anorexia/statistics.htm*). Out of a randomly chosen group of 600 U.S. women:

- **d**How many are expected to suffer from anorexia?
- **e**Find the probability that no one suffers from anorexia.

- **f**Find the probability that more than four suffer from anorexia.

Solution:

- **a** X = the number of women that suffer from anorexia
- **b**0, 1, 2, 3,... 600 (can leave off 600)
- **c** X~$\mathrm{P}(3)$
- **d**3
- **e**0.0498
- **f**0.1847

Exercise:

Problem:

The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen. (*http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html* MHLW’s Pamphlet)

- **d**Find the probability that she has no children.
- **e**Find the probability that she has fewer children than the Japanese average.
- **f**Find the probability that she has more children than the Japanese average.

Exercise:

Problem:

The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen. (*http://www.typicallyspanish.com/news/publish/article_4897.shtml*).

- **d**Find the probability that she has no children.
- **e**Find the probability that she has fewer children than the Spanish average.

- **f**Find the probability that she has more children than the Spanish average .

Solution:

- **a** X = the number of children for a Spanish woman
- **b**0, 1, 2, 3,...
- **c** $X \sim \mathrm{P}(1.47)$
- **d**0.2299
- **e**0.5679
- **f**0.4321

Exercise:

Problem:

Fertile (female) cats produce an average of 3 litters per year. *(Source: The Humane Society of the United States)*. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

- **d**No litters.
- **e**At least 2 litters.
- **f**Exactly 3 litters.

Exercise:

Problem:

A consumer looking to buy a used red Miata car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28%. We are interested in the number of dealerships she must call.

- **d**On average, how many dealerships would we expect her to have to call until she finds one that has the car?
- **e**Find the probability that she must call at most 4 dealerships.

- **f**Find the probability that she must call 3 or 4 dealerships.

Solution:

- **a** X = the number of dealers she calls until she finds one with a used red Miata
- **b**1, 2, 3,...
- **c** $X \sim \mathrm{G}(0.28)$
- **d**3.57
- **e**0.7313
- **f**0.2497

Exercise:

Problem:

Suppose that the probability that an adult in America will watch the Super Bowl is 40%. Each person is considered independent. We are interested in the number of adults in America we must survey until we find one who will watch the Super Bowl.

- **d**How many adults in America do you expect to survey until you find one who will watch the Super Bowl?
- **e**Find the probability that you must ask 7 people.
- **f**Find the probability that you must ask 3 or 4 people.

Exercise:

Problem:

A group of Martial Arts students is planning on participating in an upcoming demonstration. 6 are students of Tae Kwon Do; 7 are students of Shotokan Karate. Suppose that 8 students are randomly picked to be in the first demonstration. We are interested in the number of Shotokan Karate students in that first demonstration. Hint: Use the Hypergeometric distribution. Look in the Formulas section of 4: Discrete Distributions and in the Appendix Formulas.

- **d**How many Shotokan Karate students do we expect to be in that first demonstration?
- **e**Find the probability that 4 students of Shotokan Karate are picked for the first demonstration.
- **f**Suppose that we are interested in the Tae Kwan Do students that are picked for the first demonstration. Find the probability that all 6 students of Tae Kwan Do are picked for the first demonstration.

Solution:

- **d**4.31
- **e**0.4079
- **f**0.0163

Exercise:

Problem:

The chance of a IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20 year period. Assume each year is independent.

- **d**How many audits are expected in a 20 year period?
- **e**Find the probability that a person is not audited at all.
- **f**Find the probability that a person is audited more than twice.

Exercise:

Problem:

Refer to the previous problem. Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in 1 year. One way to solve this problem is by using the Binomial Distribution. Since n is large and p is small, another discrete distribution could be used to solve the following problems. Solve the following questions (d-f) using that distribution.

- **d**How many are expected to be audited?
- **e**Find the probability that no one was audited.
- **f**Find the probability that more than 2 were audited.

Solution:

- **d**2
- **e**0.1353
- **f**0.3233

Exercise:

Problem:

Suppose that a technology task force is being formed to study technology awareness among instructors. Assume that 10 people will be randomly chosen to be on the committee from a group of 28 volunteers, 20 who are technically proficient and 8 who are not. We are interested in the number on the committee who are **not** technically proficient.

- **d**How many instructors do you expect on the committee who are **not** technically proficient?
- **e**Find the probability that at least 5 on the committee are not technically proficient.
- **f**Find the probability that at most 3 on the committee are not technically proficient.

Exercise:

Problem:

Refer back to Exercise 4.15.12. Solve this problem again, using a different, though still acceptable, distribution.

Solution:

- **a** X = the number of seniors that participated in after-school sports all 4 years of high school
- **b**0, 1, 2, 3,... 60
- **c** $X \sim P(4.8)$
- **d**4.8
- **e**Yes
- **f**4

Exercise:

Problem:

Suppose that 9 Massachusetts athletes are scheduled to appear at a charity benefit. The 9 are randomly chosen from 8 volunteers from the Boston Celtics and 4 volunteers from the New England Patriots. We are interested in the number of Patriots picked.

- **d**Is it more likely that there will be 2 Patriots or 3 Patriots picked?

Exercise:

Problem:

On average, Pierre, an amateur chef, drops 3 pieces of egg shell into every 2 batters of cake he makes. Suppose that you buy one of his cakes.

- **d**On average, how many pieces of egg shell do you expect to be in the cake?
- **e**What is the probability that there will not be any pieces of egg shell in the cake?
- **f**Let’s say that you buy one of Pierre’s cakes each week for 6 weeks. What is the probability that there will not be any egg shell in any of the cakes?
- **g**Based upon the average given for Pierre, is it possible for there to be 7 pieces of shell in the cake? Why?

Solution:

- **a** X = the number of shell pieces in one cake
- **b**0, 1, 2, 3,...
- **c** $X \sim P(1.5)$
- **d**1.5
- **e**0.2231
- **f**0.0001
- **g**Yes

Exercise:

Problem:

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose we are interested in the number of California residents we must survey until we find a resident who does **not** have adequate earthquake supplies.

- **d**What is the probability that we must survey just 1 or 2 residents until we find a California resident who does not have adequate earthquake supplies?
- **e**What is the probability that we must survey at least 3 California residents until we find a California resident who does not have adequate earthquake supplies?
- **f**How many California residents do you expect to need to survey until you find a California resident who **does not** have adequate earthquake supplies?
- **g**How many California residents do you expect to need to survey until you find a California resident who **does** have adequate earthquake supplies?

Exercise:

Problem:

Refer to the above problem. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

- **d** What is the probability that at least 8 have adequate earthquake supplies?
- **e**Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
- **f**How many residents do you expect will have adequate earthquake supplies?

Solution:

- **d**0.0043
- **e**none
- **f**3.3

The next 2 questions refer to the following: In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages.

Exercise:

Problem:

Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked at most once.

- **d**How many pages do you expect to advertise footwear on them?
- **e**Is it probable that all 20 will advertise footwear on them? Why or why not?
- **f**What is the probability that less than 10 will advertise footwear on them?

Exercise:

Problem:

Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. This time, each page may be picked more than once.

- **d**How many pages do you expect to advertise footwear on them?
- **e**Is it probable that all 20 will advertise footwear on them? Why or why not?
- **f** What is the probability that less than 10 will advertise footwear on them?
- **g**Reminder: A page may be picked more than once. We are interested in the number of pages that we must randomly survey until we find one that has footwear advertised on it. Define the random variable X and give its distribution.
- **h**What is the probability that you only need to survey at most 3 pages in order to find one that advertises footwear on it?
- **i**How many pages do you expect to need to survey in order to find one that advertises footwear?

Solution:

- **d**3.02
- **e**No
- **f**0.9997
- **h**0.3881
- **i**6.6207 pages

Exercise:

Problem:

Suppose that you roll a fair die until each face has appeared at least once. It does not matter in what order the numbers appear. Find the expected number of rolls you must make until each face has appeared at least once.

Try these multiple choice problems.

For the next three problems: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13 year win history of 382 wins out of 1034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.
Let X = the number of games won in that upcoming month.
Exercise:

Problem: The expected number of wins for that upcoming month is:

- **A** 1.67
- **B** 12
- **C** $\frac{382}{1043}$
- **D** 4.43

Solution:

D: 4.43

Exercise:

Problem:

What is the probability that the San Jose Sharks win 6 games in that upcoming month?

- **A**0.1476
- **B**0.2336
- **C**0.7664
- **D**0.8903

Solution:

A: 0.1476

Exercise:

Problem:

What is the probability that the San Jose Sharks win at least 5 games in that upcoming month

- **A**0.3694
- **B**0.5266
- **C**0.4734
- **D**0.2305

Solution:

C: 0.4734

For the next two questions: The average number of times per week that Mrs. Plum’s cats wake her up at night because they want to play is 10. We are interested in the number of times her cats wake her up each week.

Exercise:

Problem: In words, the random variable X =

- **A** The number of times Mrs. Plum’s cats wake her up each week
- **B** The number of times Mrs. Plum’s cats wake her up each hour
- **C** The number of times Mrs. Plum’s cats wake her up each night
- **D** The number of times Mrs. Plum’s cats wake her up

Solution:

A: The number of times Mrs. Plum's cats wake her up each week

Exercise:

Problem:

Find the probability that her cats will wake her up no more than 5 times next week.

- **A**0.5000
- **B**0.9329
- **C**0.0378
- **D**0.0671

Solution:

D: 0.0671

Exercise:

Problem:

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given below. There is 5 video limit per customer at this store, so nobody ever rents more than 5 DVDs.

x	0	1	2	3	4	5
P(X=x)	0.03	0.50	0.24	?	0.07	0.04

- **A** Describe the random variable X in words.
- **B** Find the probability that a customer rents three DVDs.
- **C** Find the probability that a customer rents at least 4 DVDs.
- **D** Find the probability that a customer rents at most 2 DVDs.

Another shop, Entertainment Headquarters, rents DVDs and videogames. The probability distribution for DVD rentals per customer at this shop is given below. They also have a 5 DVD limit per customer.

x	0	1	2	3	4	5
P(X=x)	0.35	0.25	0.20	0.10	0.05	0.05

- **E** At which store is the expected number of DVDs rented per customer higher?
- **F** If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
- **G** If Video to Go expects 300 customers next week and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
- **H** Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Solution:

Partial Answer:
A: X = the number of DVDs a Video to Go customer rents
B: 0.12
C: 0.11
D: 0.77

Exercise:

Problem:

A game involves selecting a card from a deck of cards and tossing a coin. The deck has 52 cards and 12 cards are "face cards" (Jack, Queen, or King) The coin is a fair coin and is equally likely to land on Heads or Tails

- If the card is a face card and the coin lands on Heads, you win $6
- If the card is a face card and the coin lands on Tails, you win $2

- If the card is not a face card, you lose $2, no matter what the coin shows.

- **A** Find the expected value for this game (expected net gain or loss).
- **B** Explain what your calculations indicate about your long-term average profits and losses on this game.
- **C** Should you play this game to win money?

Solution:

The variable of interest is X = net gain or loss, in dollars

The face cards J, Q, K (Jack, Queen, King). There are(3)(4) = 12 face cards and 52 – 12 = 40 cards that are not face cards.

We first need to construct the probability distribution for X. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Card Event	$X net gain or loss	P(X)
Face Card and Heads	6	(12/52)(1/2) = 6/52
Face Card and Tails	2	(12/52)(1/2) = 6/52
(Not Face Card) and (H or T)	–2	(40/52)(1) = 40/52

- Expected value = (6)(6/52) + (2)(6/52) + (–2) (40/52) = –32/52
- Expected value = –$0.62, rounded to the nearest cent
- If you play this game repeatedly, over a long number of games, you would expect to lost 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

Exercise:

Problem:

You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available be sold in this lottery. In this lottery there is one $500 prize, 2 $100 prizes and 4 $25 prizes. Find your expected gain or loss.

Solution:

Start by writing the probability distribution. X is net gain or loss = prize (if any) less $10 cost of ticket

X = $ net gain or loss	P(X)
$500–$10=$490	1/100
$100–$10=$90	2/100
$25–$10=$15	4/100
$0–$10=$–10	93/100)

Expected Value = (490)(1/100) + (90)(2/100) + (15)(4/100) + (–10)(93/100) = –$2. There is an expected loss of $2 per ticket, on average.

Exercise:

Problem:

A student takes a 10 question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of **at least** 70% of the questions correct.

Solution:

- X = number of questions answered correctly
- X~B(10, 0.5)
- We are interested in AT LEAST 70% of 10 questions correct. 70% of 10 is 7. We want to find the probability that X is greater than or equal to 7. The event "at least 7" is the complement of "less than or equal to 6".
- Using your calculator's distribution menu: 1 – binomcdf(10, .5, 6) gives 0.171875
- The probability of getting at least 70% of the 10 questions correct when randomly guessing is approximately 0.172

Exercise:

Problem:

A student takes a 32 question multiple choice exam, but did not study and randomly guesses each answer. Each question has 3 possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

Solution:

- X = number of questions answered correctly
- X~B(32, 1/3)
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find P(x>24). The event "more than

24" is the complement of "less than or equal to 24".

- Using your calculator's distribution menu: 1 - binomcdf(32, 1/3, 24)
- P(x>24) = 0.00000026761
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

Exercise:

Problem:

Suppose that you are perfoming the probability experiment of rolling one fair six-sided die. Let F be the event of rolling a "4" or a "5". You are interested in how many times you need to roll the die in order to obtain the first “4 or 5” as the outcome.

- p = probability of success (event F occurs)
- q = probability of failure (event F does not occur)

- **A** Write the description of the random variable X. What are the values that X can take on? Find the values of p and q.
- **B** Find the probability that the first occurrence of event F (rolling a “4” or “5”) is on the second trial.
- **C** How many trials would you expect until you roll a “4” or “5”?

Solution:

A: X can take on the values 1, 2, 3, ... p = 2/6, q = 4/6
B: 0.2222
C: 3

**Exercises 38 - 43 contributed by Roberta Bloom

Review

This module provides a number of homework/review exercises summarizing topics related to Discrete Random Variables.

The next two questions refer to the following:

A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2000 per month. Of those respondents who charge more than $2000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

Exercise:

Problem:

What is the probability that a randomly selected respondent will spend more than $2000 AND use a credit card that gives them a mile of air travel for every dollar they charge?

- **A** $(0.30)(0.35)$
- **B** $(0.80)(0.35)$
- **C** $(0.80)(0.30)$
- **D** (0.80)

Solution:

C

Exercise:

Problem:

Based upon the above information, are using a credit card that gives a mile of air travel for each dollar spent AND charging more than $2000 per month independent events?

- **A**Yes
- **B**No, and they are not mutually exclusive either
- **C**No, but they are mutually exclusive

- **D**Not enough information given to determine the answer

Solution:

B

Exercise:

Problem:

A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women’s club and mails a questionnaire to 100 of these women selected at random. 68 questionnaires are returned. What is the population in this study?

- **A**All employed adult women
- **B**All the members of a local business and professional women’s club
- **C**The 100 women who received the questionnaire
- **D**All employed women with children

Solution:

A

The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

			Ethnic Group		
Gender	White	Asian	Hispanic	Black	American Indian
Male	400	168	115	35	16
Female	440	132	140	40	14

Exercise:

Problem: Find the probability that a student is Asian or Male.

Solution:

0.5773

Exercise:

Problem:

Find the probability that a student is Black given that the student is Female.

Solution:

0.0522

Exercise:

Problem:

A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs.
- Median = 4.5 lbs.
- Mode = 4 lbs.

- Standard deviation = 3.8 lbs.
- First quartile = 2 lbs.
- Third quartile = 8.5 lbs.

The correct statement is:

- **A**One fourth of the members lost exactly 2 pounds.
- **B**The middle fifty percent of the members lost from 2 to 8.5 lbs.
- **C**Most people lost 3.5 to 4.5 lbs.
- **D**All of the choices above are correct.

Solution:

B

Exercise:

Problem:

What does it mean when a data set has a standard deviation equal to zero?

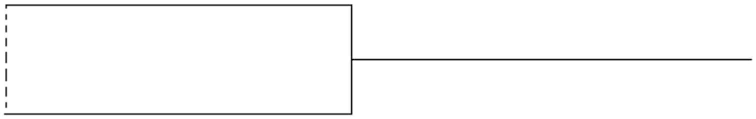
- **A**All values of the data appear with the same frequency.
- **B**The mean of the data is also zero.
- **C**All of the data have the same value.
- **D**There are no data to begin with.

Solution:

C

Exercise:

Problem: The statement that best describes the illustration below is:



- **A**The mean is equal to the median.
- **B**There is no first quartile.
- **C**The lowest data value is the median.
- **D**The median equals $\frac{(Q1+Q3)}{2}$

Solution:

C

Exercise:

Problem:

According to a recent article (San Jose Mercury News) the average number of babies born with significant hearing loss (deafness) is approximately 2 per 1000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1000 babies in an intensive care nursery.

Suppose that 1000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly 2 babies were born deaf.

Solution:

0.2709

Exercise:

Problem:

A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- 10 of the coupons are for a free gift worth \$6.
- 80 of the coupons are for a free gift worth \$8.
- 6 of the coupons are for a free gift worth \$12.
- 4 of the coupons are for a free gift worth \$40.

Based upon the financial gain or loss over the long run, should you play the game?

- **A**Yes, I expect to come out ahead in money.
- **B**No, I expect to come out behind in money.
- **C**It doesn’t matter. I expect to break even.

Solution:

B

The next four questions refer to the following: Recently, a nurse commented that when a patient calls the medical advice line claiming to have **the flu**, the chance that he/she truly has **the flu** (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have **the flu**, we are interested in how many actually have **the flu**.
Exercise:

Problem: Define the Random Variable and list its possible values.

Solution:

X = the number of patients calling in claiming to have **the flu**, who actually have **the flu**. X = 0, 1, 2, ...25

Exercise:

Problem: State the distribution of X .

Solution:

$B(25,0.04)$

Exercise:

Problem:

Find the probability that at least 4 of the 25 patients actually have **the flu**.

Solution:

0.0165

Exercise:

Problem:

On average, for every 25 patients calling in, how many do you expect to have **the flu**?

Solution:

1

The next two questions refer to the following: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

Exercise:

Problem: What kind of data was collected?

- **A**qualitative
- **B**quantitative - continuous
- **C**quantitative – discrete

Solution:

C

Exercise:

Problem: What is the population under study?

Solution:

All words used by Tom Clancy in his novels

Introduction
Continuous Random Variables: Introduction is part of the collection col10555 written by Barbara Illowsky and Susan Dean and serves as an introduction to the uniform and exponential distributions with contributions from Roberta Bloom.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Introduction

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

This chapter gives an introduction to continuous random variables and the many continuous distributions. We will be studying these continuous distributions for several chapters.

> **Note:**The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated: **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

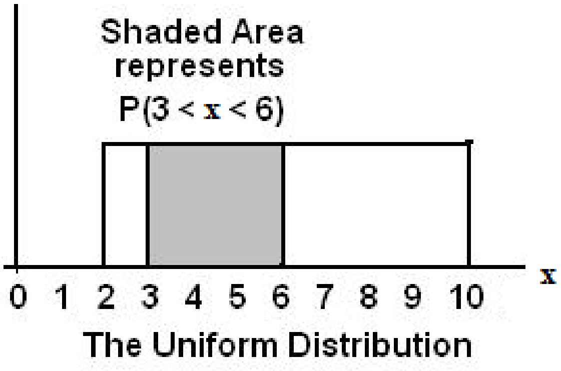
Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated: **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to 1.
- Probability is found for intervals of x values rather than for individual x values.
- $P(\mathrm{c} < x < \mathrm{d})$ is the probability that the random variable X is in the interval between the values c and d. $P(\mathrm{c} < x < \mathrm{d})$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is 0. The area below the curve, above the x-axis, and between x=c and x=c has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also 0.

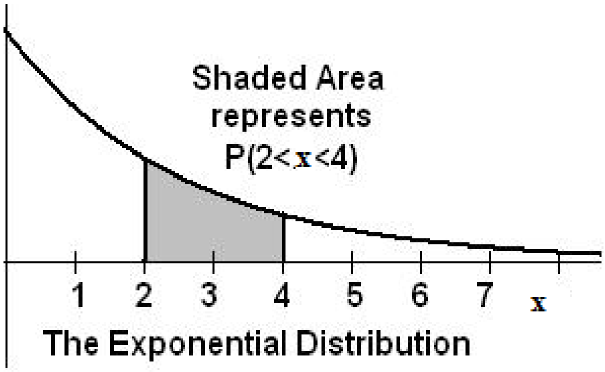
We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to best model and fit the particular situation.

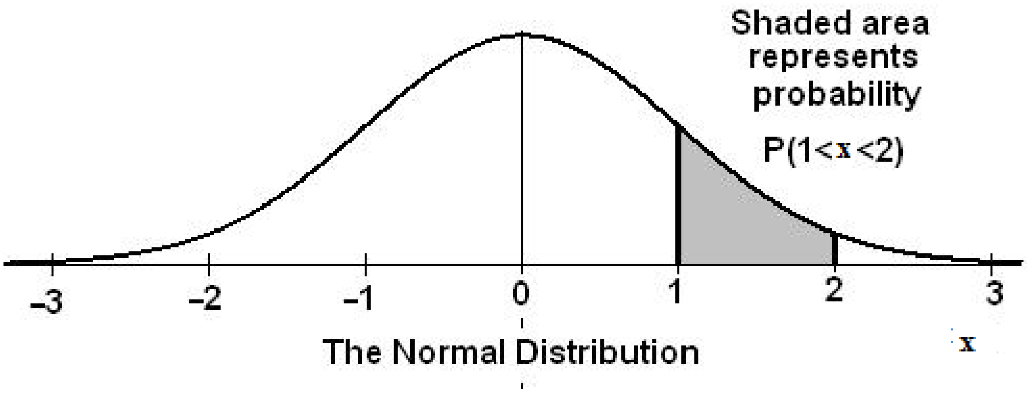
In this chapter and the next chapter, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.



The graph shows a Uniform Distribution with the area between x=3 and x=6 shaded to represent the probability that the value of the random variable X is in the interval between 3 and 6.



The graph shows an Exponential Distribution with the area between x=2 and x=4 shaded to represent the probability that the value of the random variable X is in the interval between 2 and 4.



The graph shows the Standard Normal Distribution with the area between x=1 and x=2 shaded to represent the probability that the value of the random variable X is in the interval between 1 and 2.

**With contributions from Roberta Bloom

Glossary

Uniform Distribution

A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: X~$U(a,b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(X) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Exponential Distribution

A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: X~$\text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = \text{me}^{-\text{mx}}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-\text{mx}}$.

Continuous Probability Functions

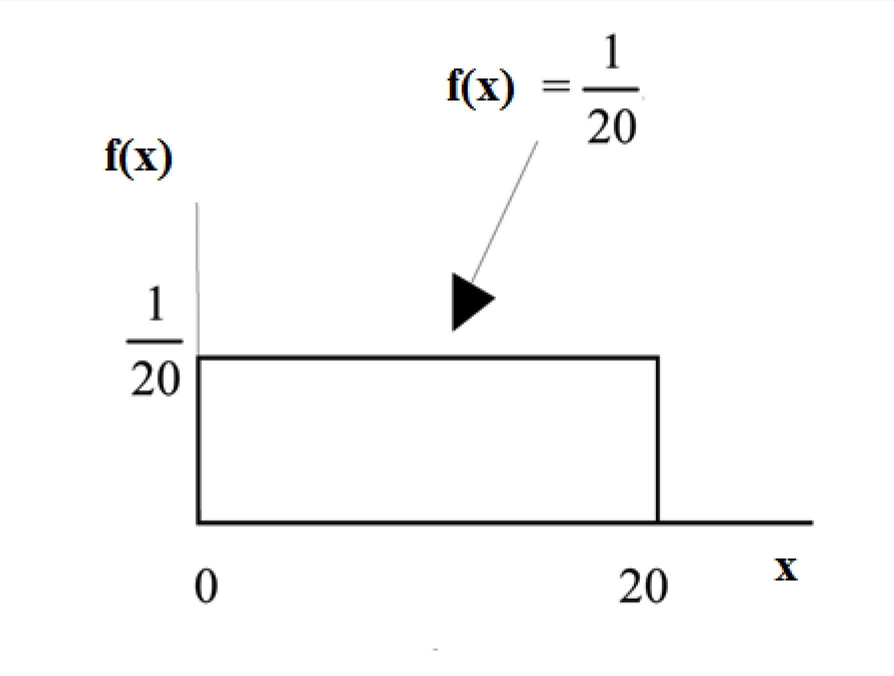
This module introduces the continuous probability function and explores the relationship between the probability of X and the area under the curve of f(X).

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one.

For continuous probability distributions, PROBABILITY = AREA.

Example:

Consider the function $f(x)=\frac{1}{20}$ for $0 \leq x \leq 20$. x = a real number. The graph of $f(x)=\frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x=0$ and $x=20$, inclusive .



$f(x) = \frac{1}{20}$ **for** $0\leq\ x\ \leq 20.$

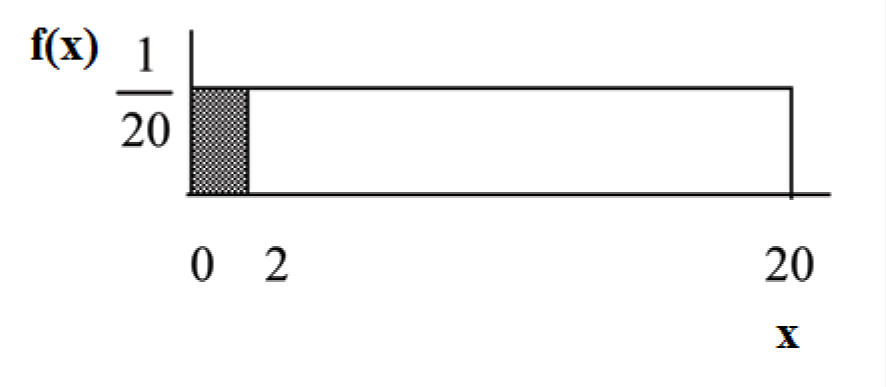
The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20.$

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x-axis is the area of a rectangle with base = 20 and height =$\frac{1}{20}$.

$\text{AREA} = 20 \cdot \frac{1}{20} = 1$

This particular function, where we have restricted x so that the area between the function and the x-axis is 1, is an example of a continuous probability density function. It is used as a tool to calculate probabilities.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ **and the x-axis where** $0 < x < 2$ **.**



$\text{AREA} = (2 - 0) \cdot \frac{1}{20} = 0.1$

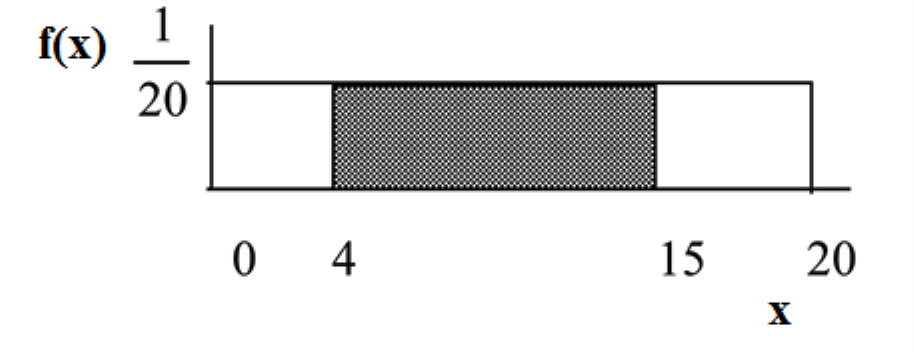
$(2 - 0) = 2 = \text{base of a rectangle}$

$\frac{1}{20}$ = the height.

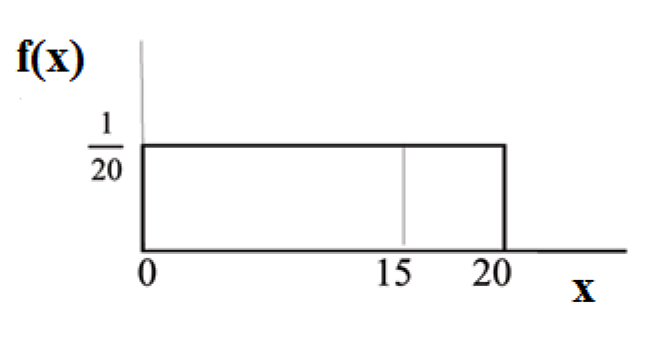
The area corresponds to a probability. The probability that x is between 0 and 2 is 0.1, which can be written mathematically as

$\text{P(0<x<2)} = \text{P(x<2)} = 0.1.$

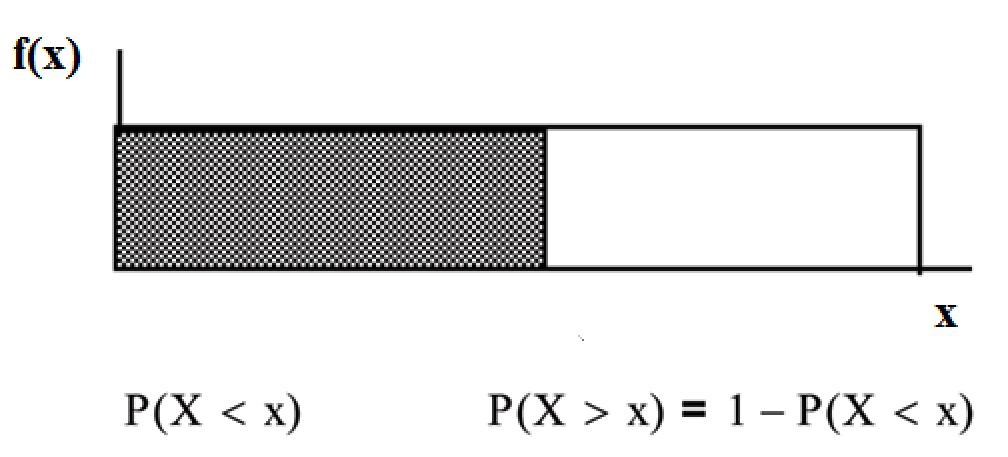
Suppose we want to find the area between $f(x) = \frac{1}{20}$ **and the x-axis where** $4 < x < 15$ **.**



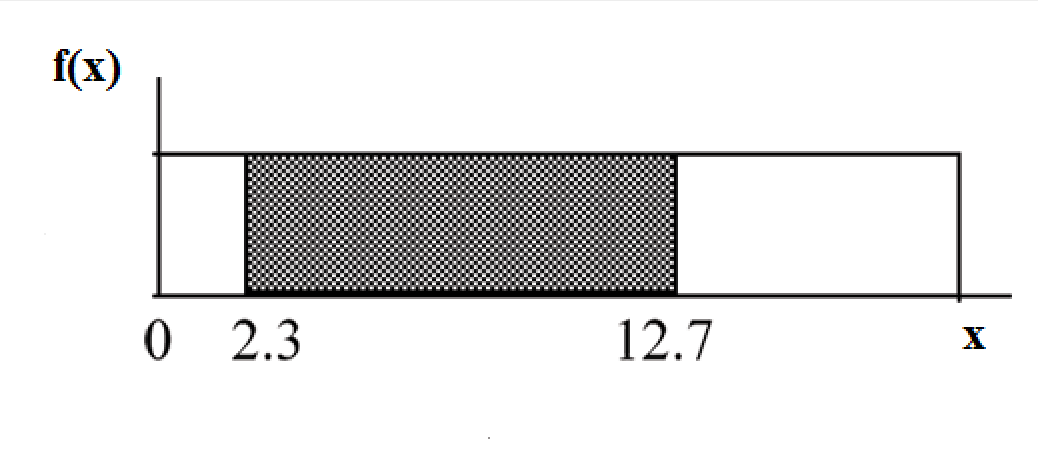
$\text{AREA} = (15 - 4) \cdot \frac{1}{20} = 0.55$
$(15 - 4) = 11 = \text{the base of a rectangle}$
$\frac{1}{20}$ = the height.
The area corresponds to the probability $P(4 < x < 15) = 0.55$.
Suppose we want to find $P(x=15)$**.** On an x-y graph, $x=15$ is a vertical line. A vertical line has no width (or 0 width). Therefore,
$P(x=15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$.



$P(X \leq x)$ (can be written as $P(X < x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows:
$P(X > x) = 1 - P(X < x)$.



Label the graph with f(x) and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.



$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = \left(12.7 - 2.3\right)\left(\frac{1}{20}\right) = 0.52$

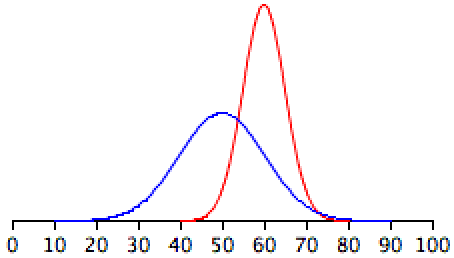
Introduction to Normal Distributions

Normal distributions are commonly used in Statistics. While normal distributions can be quite different, they can all be represented mathematically and they all have distinct features that will be discussed in this chapter.

The normal distribution is the most important and widely used distribution in statistics. It is sometimes called the **bell curve** although the tonal qualities of such a bell would be less than pleasing. It is also called the **Gaussian curve** after the mathematician Karl-Friedrich Gauss.

Strictly speaking, it is not correct to talk about **the normal distribution** since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. [link] shows two normal distributions. The blue distribution has a mean of 50 and a standard deviation of 10; the distribution in red has a mean of 60 and a standard deviation of 5. Both distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

Varieties of Normal Distributions



Normal distributions differing in mean and standard deviation.

The density of the normal distribution (the height for a given value on the x axis) of the normal distribution is shown below ([link]). The parameters μ and σ are the mean and standard deviation repectively and define the normal distribution. The symbol e is the base of the natural logarithm and π is the constant pi.

Equation:

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will **not** be referring back to it in later sections.

Some features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean
7. 95% of the area of a normal distribution is within two standard deviations of the mean
8. 99.7% of the area of a normal distribution is within three standard deviations of the mean

The Standard Normal Distribution

This module introduces standard normal distribution and standardizing the distribution.

As previously discussed, normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution** and is typically represented by Z.

Areas underneath the normal distribution are often represented by tables of the standard normal distribution. A portion of a table of the standard normal distribution is shown in [link].

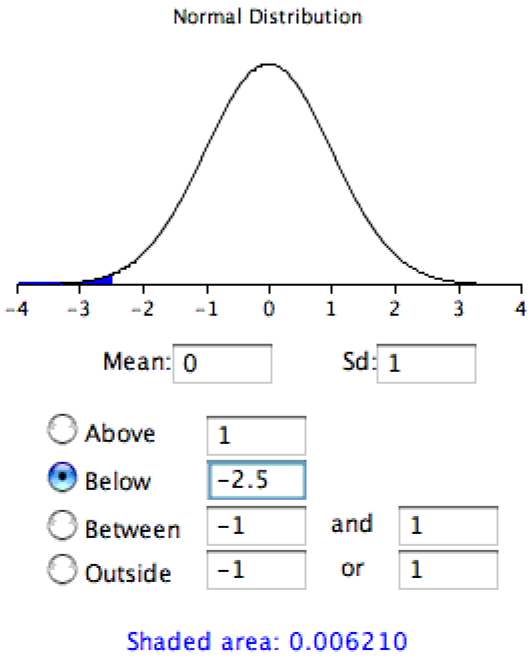
Z	**Area left of Z**
-2.50	0.0062
-2.49	0.0064
-2.48	0.0066
-2.47	0.0068
-2.46	0.0069
-2.45	0.0071
-2.44	0.0073
-2.43	0.0075
-2.42	0.0078

Z	**Area left of Z**
-2.41	0.0080
-2.40	0.0082
-2.39	0.0084
-2.38	0.0087
-2.37	0.0089
-2.36	0.0091
-2.35	0.0094
-2.34	0.0096
-2.33	0.0099
-2.32	0.0102

A portion of a table of the standard normal distribution.

The first column titled Z contains values of the standard normal random variable; the second column contains the area below the curve to the left of z. Since the distribution has a mean of 0 and a standard deviation of 1, the Z column is equal to the number of standard deviations below (or above) the mean. For example, a z of -2.5 represents a value 2.5 standard deviations below the mean. The area below the curve to the left of z=-2.5 is 0.0062.

The same information can be obtained using a calculator or the following Java applet. [link] shows how it can be used to compute the area below the standard normal curve to the left of -2.5. Note that the mean is set to 0 and the standard deviation is set to 1.



Mean: 0 Sd: 1

○ Above 1

● Below -2.5

○ Between -1 and 1

○ Outside -1 or 1

Shaded area: 0.006210

An example from the applet.

Note: Calculate Areas

A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

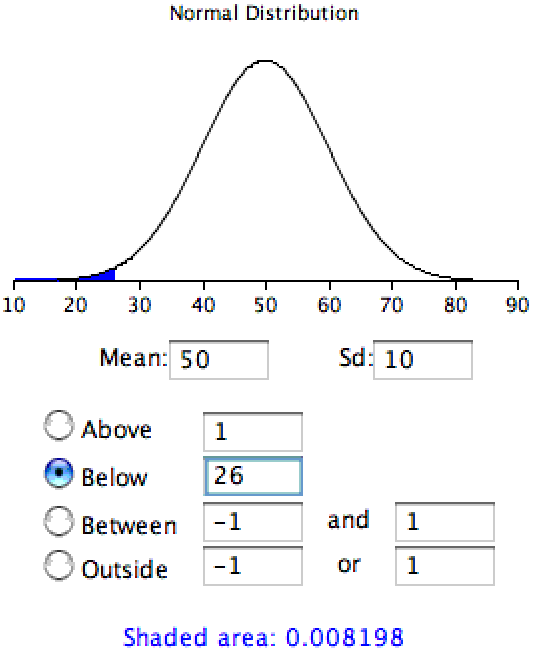
$$z = \frac{x - \mu}{\sigma}$$

where z is the value on the standard normal distribution, x is the value on the original distribution, μ is the mean of the original distribution and σ is the standard deviation of the original distribution.

As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26? Applying the formula we obtain

$$z=\frac{25-50}{10}=-2.4$$

From [link], we can see that 0.0082 of the distribution is below -2.4. There is no need to transform to Z if you are using a technology or the applet shown in [link].



Mean: 50 Sd: 10

○ Above 1

● Below 26

○ Between -1 and 1

○ Outside -1 or 1

Shaded area: 0.008198

Area below 36 in a normal distribution with a mean of 50 and a standard deviation of 10.

If all the values in a distribution are transformed to z-scores, then the distribution will have a mean of 0 and a standard distribution. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called **standardizing** the distribution.

Calculate Areas

Applet failed to run. No Java plug-in was found.

Glossary

standard normal distribution

The standard normal distribution is a **normal distribution** with a mean of 0 and a standard deviation of 1.

Z-scores

If X is a normally distributed random variable and X~N(μ, σ), then the z-score is:

Equation:

$$z=\frac{x-\mu}{\sigma}$$

The z-score tells you how many standard deviations that the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of 0.

Example:

Suppose X ~ N(5, 6). This says that X is a normally distributed random variable with mean μ = 5 and standard deviation σ = 6. Suppose x = 17. Then:

Equation:

$$z=\frac{x-\mu}{\sigma}=\frac{17-5}{6}=2$$

This means that x = 17 is **2 standard deviations** (2σ) above or to the right of the mean μ = 5. The standard deviation is σ = 6.

Notice that:

Equation:

$$5+2\cdot 6=17 \qquad \text{(The pattern is } \mu+z\sigma=x.\text{)}$$

Now suppose x=1. Then:

Equation:

$$z=\frac{x-\mu}{\sigma}=\frac{1-5}{6}=-0.67 \qquad \text{(rounded to two decimal places)}$$

This means that $\mathrm{x} = 1$ **is 0.67 standard deviations** $(\text{- } 0.67\sigma)$ **below or to the left of the mean** $\mu = 5$**. Notice that:**
$5 + (-0.67)(6)$ is approximately equal to 1 (This has the pattern $\mu + (-0.67)\sigma = 1$)
Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ.

Example:
Some doctors believe that a person can lose 5 pounds, on the average, in a month by reducing his/her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of 2 pounds. X~N$(5, 2)$. Fill in the blanks.

Exercise:

Problem:

Suppose a person **lost** 10 pounds in a month. The z-score when $\mathrm{x} = 10$ pounds is $\mathrm{z} = 2.5$ (verify). This z-score tells you that $\mathrm{x} = 10$ is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).

Solution:

This z-score tells you that $\mathrm{x} = 10$ is **2.5** standard deviations to the **right** of the mean **5**.

Exercise:

Problem:

Suppose a person **gained** 3 pounds (a negative weight loss). Then z = __________. This z-score tells you that $\mathrm{x} = \text{-}3$ is ________ standard deviations to the __________ (right or left) of the mean.

Solution:

z = **-4**. This z-score tells you that x = -3 is **4** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: X ~N$(5, 6)$ and Y ~ N$(2, 1)$. If x = 17, then $z = 2$. (This was previously shown.) If y = 4, what is z?

Equation:

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2 \qquad \text{where } \mu{=}2 \text{ and } \sigma{=}1.$$

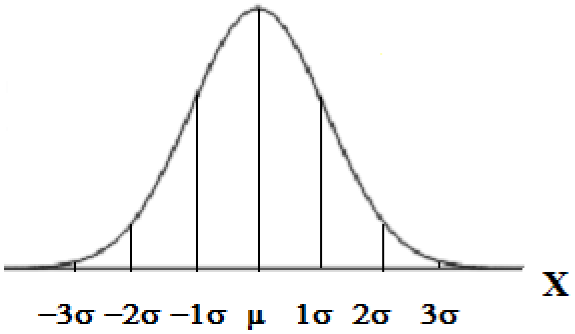
The z-score for y = 4 is z = 2. This means that 4 is z = 2 standard deviations to the right of the mean. Therefore, x = 17 and y = 4 are both 2 (of **their**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose X ~N$(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a 6 week period and Y ~N$(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since x = 17 and y = 4 are each 2 standard deviations to the right of their means, they represent the same weight gain **relative to their means**.

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ then the **Empirical Rule** says (See the figure below)

- About 68.27% of the x values lie between -1σ and +1σ of the mean μ (within 1 standard deviation of the mean).
- About 95.45% of the x values lie between -2σ and +2σ of the mean μ (within 2 standard deviations of the mean).
- About 99.73% of the x values lie between -3σ and +3σ of the mean μ (within 3 standard deviations of the mean). Notice that almost all the x values lie within 3 standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and -1, respectively.
- The z-scores for +2σ and –2σ are +2 and -2, respectively.
- The z-scores for +3σ and –3σ are +3 and -3 respectively.



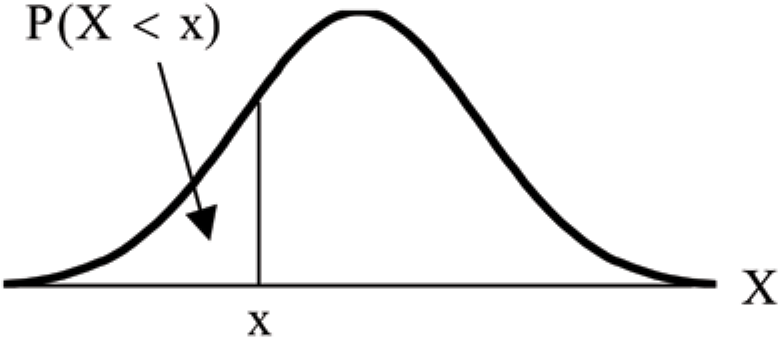
The Empirical Rule is also known as the 68-95-99.7 Rule.

Example:
Suppose X has a normal distribution with mean 50 and standard deviation 6.

- About 68.27% of the x values lie between -1σ = (-1)(6) = -6 and 1σ = (1)(6) = 6 of the mean 50. The values 50 - 6 = 44 and 50 + 6 = 56 are within 1 standard deviation of the mean 50. The z-scores are -1 and +1 for 44 and 56, respectively.
- About 95.45% of the x values lie between -2σ = (-2)(6) = -12 and 2σ = (2)(6) = 12 of the mean 50. The values 50 - 12 = 38 and 50 + 12 = 62 are within 2 standard deviations of the mean 50. The z-scores are -2 and 2 for 38 and 62, respectively.
- About 99.73% of the x values lie between -3σ = (-3)(6) = -18 and 3σ = (3)(6) = 18 of the mean 50. The values 50 - 18 = 32 and 50 + 18 = 68 are within 3 standard deviations of the mean 50. The z-scores are -3 and +3 for 32 and 68, respectively.

Normal Distribution: Areas to the Left and Right of x

The arrow in the graph below points to the area to the left of x. This area is represented by the probability $P(X < x)$**.** Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.



The area to the right is then $P(X > x) = 1 - P(X < x)$**.**

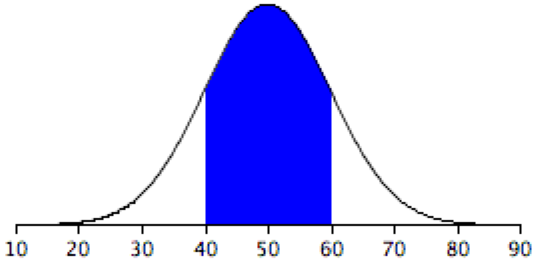
Remember, $P(X < x) =$ **Area to the left** of the vertical line through x.

$P(X > x) = 1 - P(X < x) =$. **Area to the right** of the vertical line through x

$P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

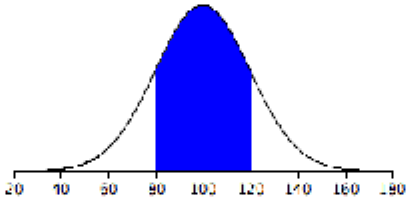
Areas of Normal Distributions

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. [link] shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.



Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

[link] shows a normal distribution with a mean of 100 and a standard deviation of 20. As in Figure 1, 68% of the distribution is within one standard deviation of the mean.

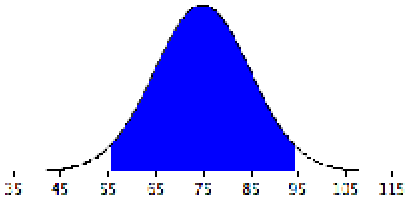


Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within one standard deviation (20) of the mean (100).

The normal distributions shown in [link] and [link] are specific examples of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

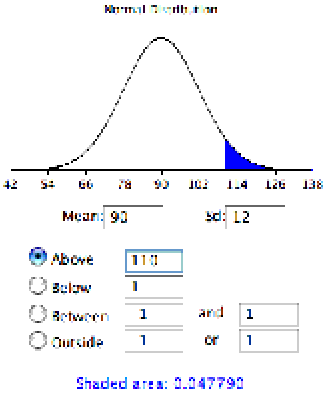
Note: 68% of the area of any normal distribution is within one standard deviation of the mean

[link] shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.



A normal distribution with a mean of 75 and a standard deviation of 10. 95% of the area is within 1.96 standard deviations of the mean.

The Java applet "Calculate Area for a given X " can be used to calculate areas under the normal distribution. Use it to find the proportion of a normal distribution with a mean of 90 and a standard deviation of 12 that is above 110. Set the mean to 90 and the standard deviation to 12. Then enter "110" in the box to the right of the radio button "Above." At the bottom of the display you will see that the shaded area is 0.04779. See if you can use the applet to find that the area between 115 and 120 is 0.012401.



Mean: 90 Sd: 12

● Above 110

○ Below 1

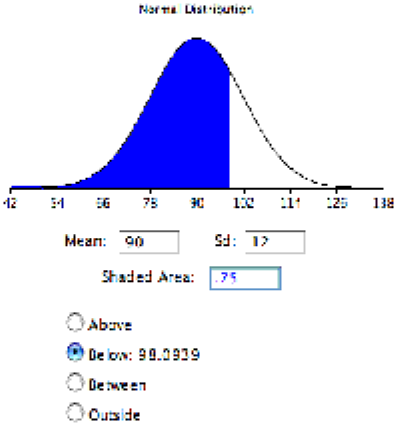
○ Between 1 and 1

○ Outside 1 or 1

Shaded area: 0.047790

Display from applet showing the area above 110.

The applet "Calculate X for a given Area" works in reverse. For example, say you wanted to find the score corresponding to the 75th percentile of a normal distribution with a mean of 90 and a standard deviation of 12. You enter 90 for the mean and 12 for the standard deviation. Then, enter 0.75 for the shaded area and click the "Below" button. The area below 98.0939 is 0.75.



Mean: 90 Sd: 12

Shaded Area: .75

○ Above

● Below: 98.0939

○ Between

○ Outside

Display from applet showing that the 75th percentile is 98.093.

Calculations of Probabilities

Probabilities are calculated by using technology. There are instructions in the chapter for the TI-83+ and TI-84 calculators.

Note: In the Table of Contents for **Collaborative Statistics**, entry **15. Tables** has a link to a table of normal probabilities. Use the probability tables if so desired, instead of a calculator. The tables include instructions for how to use then.

Example:
If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Example:
The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.
Exercise:

Problem:

Find the probability that a randomly selected student scored more than 65 on the exam.

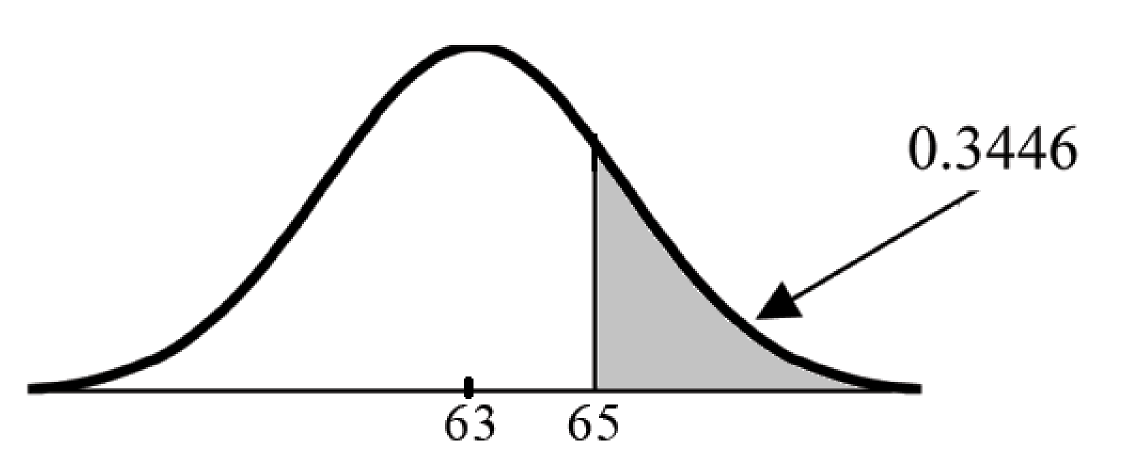
Solution:

Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$ (calculator or computer)



The probability that one student scores more than 65 is 0.3446.

Using the TI-83+ or the TI-84 calculators, the calculation is as follows. Go into `2nd DISTR`.

After pressing `2nd DISTR`, press `2:normalcdf`.

The syntax for the instructions are shown below.

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1`, the `EE` key (a 2nd key) and then `99`. Or, you can enter `10^99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be -1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Note: The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$z = \frac{65-63}{5} = 0.4$. Area to the left is 0.6554.
$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$

Exercise:

Problem:

Find the probability that a randomly selected student scored less than 85.

Solution:

Draw a graph.

Then find $P(x < 85)$. Shade the graph. $P(x < 85) = 1$ (calculator or computer)

The probability that one student scores less than 85 is approximately 1 (or 100%).

The TI-instructions and answer are as follows:

normalcdf(0,85,63,5) = 1 (rounds to 1)

Exercise:

Problem:

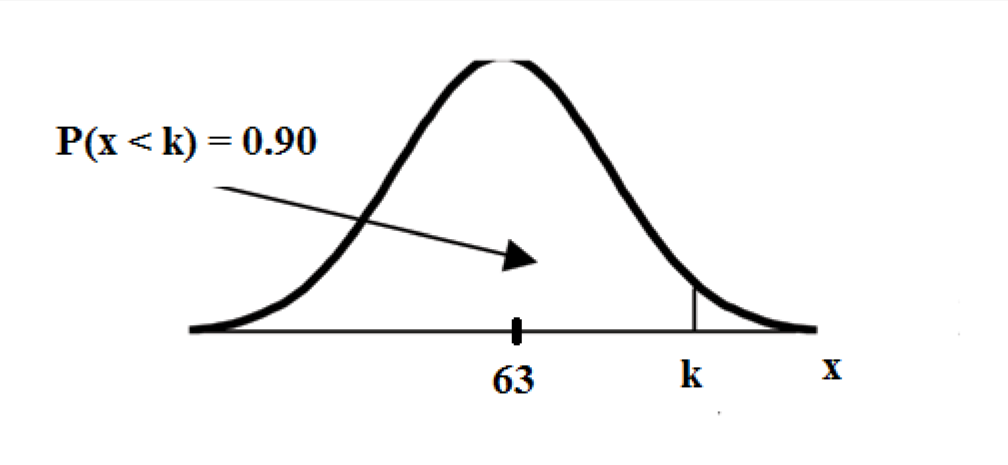
Find the 90th percentile (that is, find the score k that has 90 % of the scores below k and 10% of the scores above k).

Solution:

Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k and 10% are the same or higher. k is often called a **critical value**.

$k = 69.4$ (calculator or computer)



The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. For the TI-83+ or TI-84 calculators, use `invNorm` in `2nd DISTR`. invNorm(area to the left, mean, standard deviation) For this problem, invNorm(0.90,63,5) = 69.4

Exercise:

Problem:

Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution:

Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

invNorm(0.70,63,5) = 65.6

Example:
A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

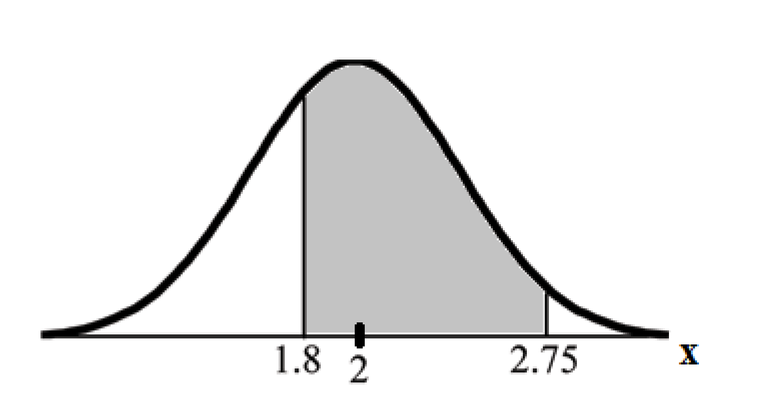
Find the probability that a household personal computer is used between 1.8 and 2.75 hours per day.

Solution:

Let X = the amount of time (in hours) a household personal computer is used for entertainment. $x \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$



normalcdf(1.8,2.75,2,0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

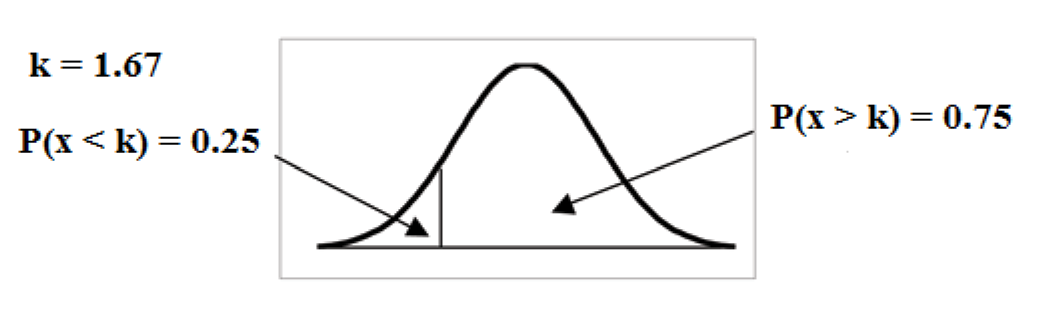
Exercise:

Problem:

Find the maximum number of hours per day that the bottom quartile of households use a personal computer for entertainment.

Solution:

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.



invNorm(0.25,2,.5) = 1.66

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Summary of Formulas

Formula

Normal Probability Distribution

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Formula

Standard Normal Probability Distribution

$Z \sim N(0, 1)$

z = a standardized value (z-score)

mean = 0 standard deviation = 1

Formula

Finding the kth Percentile

To find the **kth** percentile when the z-score is known: $k = \mu + (z)\sigma$

Formula

z-score

$z = \frac{x - \mu}{\sigma}$

Formula

Finding the area to the left

The area to the left: $P(X < x)$

Formula

Finding the area to the right

The area to the right: $P(X > x) = 1 - P(X < x)$

Practice: The Normal Distribution

Student Learning Outcomes

- The student will analyze data following a normal distribution.

Given

The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for 3 years. We are interested in the length of time a CD player lasts.

Normal Distribution

Exercise:

Problem: Define the Random Variable X in words. $X =$

Exercise:

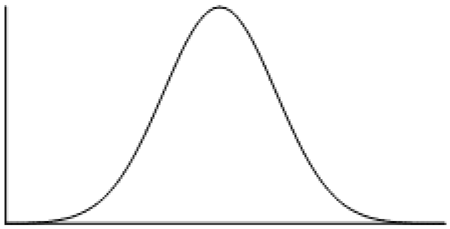
Problem: X~

Exercise:

Problem:

Find the probability that a CD player will break down during the guarantee period.

- **a** Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



- **b** $P(0 < x < \_\_\_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_\_\_\_$ (Use zero (0) for the minimum value of x.)

Solution:

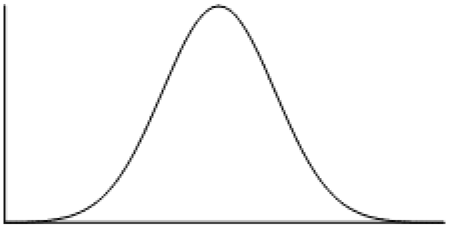
- **b** 3,0.1979

Exercise:

Problem:

Find the probability that a CD player will last between 2.8 and 6 years.

- **a** Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



- **b** $P(\_\_\_\_\_\_\_\_ < x < \_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_\_\_\_$

Solution:

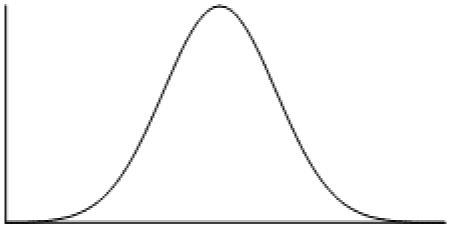
- **b** 2.8,6,0.7694

Exercise:

Problem:

Find the 70th percentile of the distribution for the time a CD player lasts.

- **a** Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.



- **b** $P(x < k) =$ __________. Therefore, $k =$ ___________.

Solution:

- **b** 0.70,4.78years

Homework

Exercise:

Problem:

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X =height of the individual.

- **a** X~_______(_______,_______)
- **b** Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph and write a probability statement.
- **c** Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
- **d** The middle 40% of heights fall between what two values? Sketch the graph and write the probability statement.

Solution:

- **a** $N(66,2.5)$
- **b**0.5404
- **c**No
- **d**Between 64.7 and 67.3 inches

Exercise:

Problem:

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X =IQ of an individual.

- **a** X~_______(_______,_______)
- **b** Find the probability that the person has an IQ greater than 120. Include a sketch of the graph and write a probability statement.

- **c** Mensa is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the Mensa organization. Sketch the graph and write the probability statement.
- **d** The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let $X =$percent of fat calories.

- **a** X~_______(_______,_______)
- **b** Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
- **c** Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

Solution:

- **a** $N(36,10)$
- **b** 0.3446
- **c** 29.3

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

- **a** If $X =$ distance in feet for a fly ball, then X~ _______(_______,_______)
- **b** If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
- **c** Find the 80th percentile of the distribution of fly balls. Sketch the graph and write the probability statement.

Exercise:

Problem:

In China, 4-year-olds average 3 hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly survey one Chinese 4-year-old living in a rural area. We are interested in the amount of time the child spends alone per day. (Source: **San Jose Mercury News**)

- **a** In words, define the random variable X. $X =$
- **b** X~
- **c** Find the probability that the child spends less than 1 hour per day unsupervised. Sketch the graph and write the probability statement.
- **d** What percent of the children spend over 10 hours per day unsupervised?
- **e** 70% of the children spend at least how long per day unsupervised?

Solution:

- **a** the time (in hours) a 4-year-old in China spends unsupervised per day
- **b** $N(3{,}1.5)$
- **c** 0.0912
- **d** 0

- **e** 2.21 hours

Exercise:

Problem:

In the 1992 presidential election, Alaska’s 40 election districts averaged 1956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let $X =$ number of votes for President Clinton for an election district. (Source: **The World Almanac and Book of Facts**)

- **a** State the approximate distribution of X. X~
- **b** Is 1956.8 a population mean or a sample mean? How do you know?
- **c** Find the probability that a randomly selected district had fewer than 1600 votes for President Clinton. Sketch the graph and write the probability statement.
- **d** Find the probability that a randomly selected district had between 1800 and 2000 votes for President Clinton.
- **e** Find the third quartile for votes for President Clinton.

Exercise:

Problem:

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of 7 days.

- **a** In words, define the random variable X. $X =$
- **b** X~
- **c** If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
- **d** 60% of all of these types of trials are completed within how many days?

Solution:

- **a**The duration of a criminal trial
- **b** $N(21,7)$
- **c** 0.3341
- **d** 22.77

Exercise:

Problem:

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a 7 lap race) with a standard deviation of 2.28 seconds . The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps. (Source: log book of Terri Vogel)

- **a** In words, define the random variable X. $X =$
- **b** X~
- **c** Find the percent of her laps that are completed in less than 130 seconds.
- **d** The fastest 3% of her laps are under ________ .
- **e** The middle 80% of her laps are from ________ seconds to ________ seconds.

Exercise:

Problem:

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X =time in line. Below are the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

- **a** Calculate the sample mean and the sample standard deviation.
- **b** Construct a histogram. Start the $x - \mathrm{axis}$ at -0.375 and make bar widths of 2 minutes.
- **c** Draw a smooth curve through the midpoints of the tops of the bars.
- **d** In words, describe the shape of your histogram and smooth curve.
- **e** Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by X~
- **f** Use the distribution in (e) to calculate the probability that a person will wait fewer than 6.1 minutes.
- **g** Determine the cumulative relative frequency for waiting less than 6.1 minutes.
- **h** Why aren’t the answers to (f) and (g) exactly the same?
- **i** Why are the answers to (f) and (g) as close as they are?

- **j** If only 10 customers were surveyed instead of 50, do you think the answers to (f) and (g) would have been closer together or farther apart? Explain your conclusion.

Solution:

- **a** The sample mean is 5.51 and the sample standard deviation is 2.15
- **e** $N(5.51, 2.15)$
- **f** 0.6081
- **g** 0.64

Exercise:

Problem:

Suppose that Ricardo and Anita attend different colleges. Ricardo’s GPA is the same as the average GPA at his school. Anita’s GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

- **a** Ricardo’s actual GPA is lower than Anita’s actual GPA.
- **b** Ricardo is not passing since his z-score is zero.
- **c** Anita is in the 70th percentile of students at her college.

Exercise:

Problem:

Below is a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse racing or motor racing stadiums. (Source: **http://en.wikipedia.org/wiki/List_of_stadiums_by_capacity**)

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

- **a** Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
- **b** Construct a histogram of the data.
- **c** Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
- **d** In words, describe the shape of your histogram and smooth curve.
- **e** Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by X~
- **f** Use the distribution in (e) to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
- **g** Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a

maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
- **h** Why aren’t the answers to (f) and (g) exactly the same?

Solution:

- **a** The sample mean is 60,136.4 and the sample standard deviation is 10,468.1.
- **e** $N(60136.4{,}10468.1)$
- **f** 0.7440
- **g** 0.7167

Try These Multiple Choice Questions

The questions below refer to the following: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem: What is the median recovery time?

- **A** 2.7
- **B** 5.3
- **C** 7.4
- **D** 2.1

Solution:

B

Exercise:

Problem:

What is the z-score for a patient who takes 10 days to recover?

- **A** 1.5
- **B** 0.2
- **C** 2.2
- **D** 7.3

Solution:

C

Exercise:

Problem:

What is the probability of spending more than 2 days in recovery?

- **A** 0.0580
- **B** 0.8447
- **C** 0.0553
- **D** 0.9420

Solution:

D

Exercise:

Problem: The 90th percentile for recovery times is?

- **A** 8.89
- **B** 7.07
- **C** 7.99
- **D** 4.32

Solution:

C

The questions below refer to the following: The length of time to find a parking space at 9 A.M. follows a normal distribution with a mean of 5 minutes and a standard deviation of 2 minutes.

Exercise:

Problem:

Based upon the above information and numerically justified, would you be surprised if it took less than 1 minute to find a parking space?

- **A** Yes
- **B** No
- **C** Unable to determine

Solution:

A

Exercise:

Problem:

Find the probability that it takes at least 8 minutes to find a parking space.

- **A** 0.0001
- **B** 0.9270
- **C** 0.1862
- **D** 0.0668

Solution:

D

Exercise:

Problem:

Seventy percent of the time, it takes more than how many minutes to find a parking space?

- **A** 1.24
- **B** 2.41
- **C** 3.95
- **D** 6.05

Solution:

C

Exercise:

Problem:

If the mean is significantly greater than the standard deviation, which of the following statements is true?

- **I** The data cannot follow the uniform distribution.
- **II** The data cannot follow the exponential distribution..
- **III** The data cannot follow the normal distribution.

- **A** I only
- **B** II only
- **C** III only
- **D** I, II, and III

Solution:

B

Review

The next two questions refer to: $X \sim U(3, 13)$
Exercise:

Problem: Explain which of the following are false and which are true.

- **a** $f(x) = \frac{1}{10}, 3 \leq x \leq 13$
- **b** There is no mode.
- **c** The median is less than the mean.
- **d** $P(x > 10) = P(x \leq 6)$

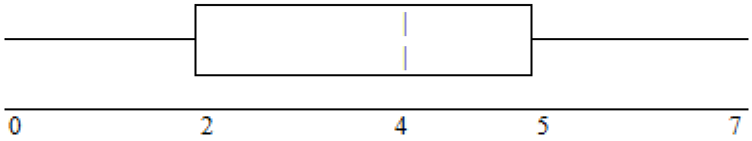
Solution:

- **a** True
- **b** True
- **c** False – the median and the mean are the same for this symmetric distribution
- **d** True

Exercise:

Problem: Calculate:

- **a** Mean
- **b** Median
- **c** 65th percentile.



Solution:

- **a** 8

- **b** 8
- **c** $P(x < k) = 0.65 = (k - 3) * (\frac{1}{10})$. $k = 9.5$

Exercise:

Problem: Which of the following is true for the above box plot?

- **a** 25% of the data are at most 5.
- **b** There is about the same amount of data from 4 – 5 as there is from 5 – 7.
- **c** There are no data values of 3.
- **d** 50% of the data are 4.

Solution:

- **a** False – $\frac{3}{4}$ of the data are at most 5
- **b** True – each quartile has 25% of the data
- **c** False – that is unknown
- **d** False – 50% of the data are 4 or less

Exercise:

Problem:

If $P(G \mid H) = P(G)$, then which of the following is correct?

- **A** G and H are mutually exclusive events.
- **B** $P(G) = P(H)$
- **C** Knowing that H has occurred will affect the chance that G will happen.
- **D** G and H are independent events.

Solution:

D

Exercise:

Problem:

If $P(J) = 0.3$, $P(K) = 0.6$, and J and K are independent events, then explain which are correct and which are incorrect.

- **A** $P(J \text{ and } K) = 0$
- **B** $P(J \text{ or } K) = 0.9$
- **C** $P(J \text{ or } K) = 0.72$
- **D** $P(J) \neq P(J \mid K)$

Solution:

- **A** False - J and K are independent so they are not mutually exclusive which would imply dependency (meaning P(J and K) is not 0).
- **B** False - see answer C.
- **C** True - P(J or K) = P(J) + P(K) - P(J and K) = P(J) + P(K) - P(J)P(K) = 0.3 + 0.6 - (0.3)(0.6) = 0.72. Note that P(J and K) = P(J)P(K) because J and K are independent.
- **D** False - J and K are independent so P(J) = P(J|K).

Exercise:

Problem:

On average, 5 students from each high school class get full scholarships to 4-year colleges. Assume that most high school classes have about 500 students.

X = the number of students from a high school class that get full scholarships to 4-year school. Which of the following is the distribution of X?

- **A**P(5)
- **B**B(500,5)
- **C**Exp(1/5)

- **D**N(5, (0.01)(0.99)/500)

Solution:

A

Introduction
This module provides a brief introduction to the Central Limit Theorem.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Means.
- Apply and interpret the Central Limit Theorem for Sums.

Introduction

Why are we so concerned with means? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study means and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n and n is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, n, that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more observations are needed for the sample means or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times.

Each time a person rolls more than one die, he/she calculates the sample **mean** of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5} = 3.4.$ The 3.4 is one mean when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more means for a total of 10 means.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "mean" when you roll one die, is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using 2 dice. Do the sample means show any kind of pattern?

Draw the graph for the means using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the means using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\bar{X}\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(\mathrm{n\mu}, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

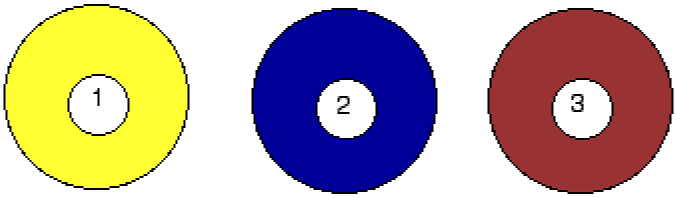
Introduction to Sampling Distributions

Suppose you randomly sampled 10 people from the population of women in Houston Texas between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall that **inferential statistics** concerns generalizing from a **sample** to a **population**. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population **parameter**. (In this example, the sample means are sample statistics and the sample parameter is the population mean.) As the later portions of this chapter show (Sampling Distribution of the Mean and Sampling Distribution of Difference Between Means), these determinations are based on **sampling distributions**.

Discrete Distributions

We will illustrate the concept of sampling distributions with a simple example. [link] shows three pool balls, each with a number on it. Two of the balls are selected randomly (with replacement) and the average of their numbers is computed.



The pool balls.

All possible outcomes are shown in [link].

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

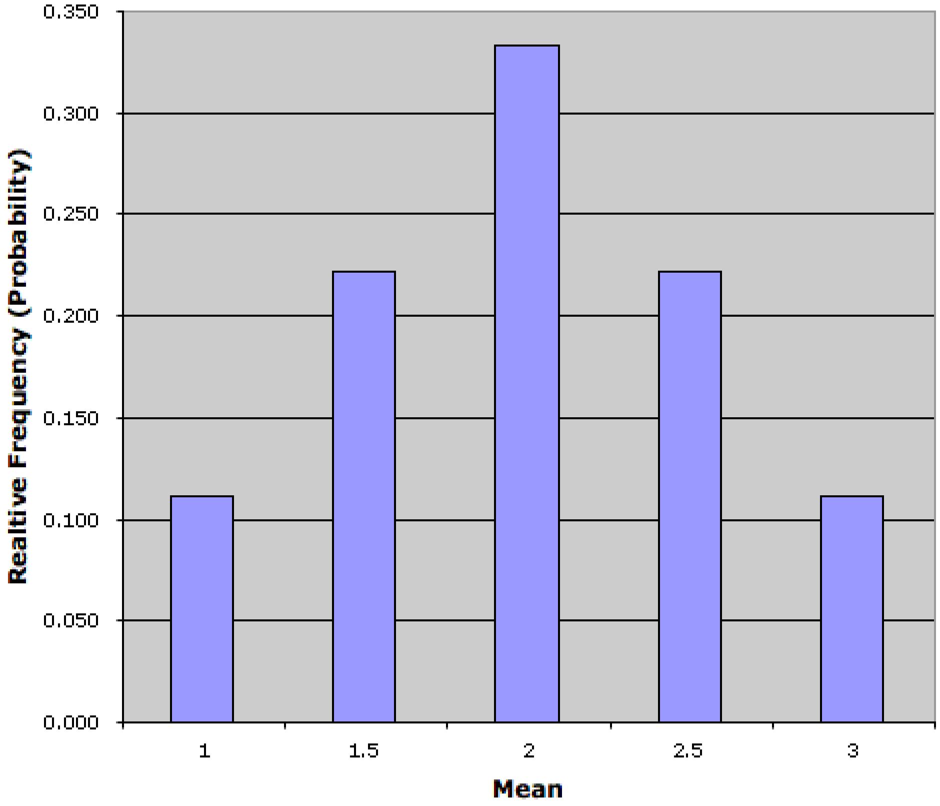
All possible outcomes when two balls are sampled

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in [link]. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Mean	Frequency	Relative Frequency
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

Frequencies of means for N = 2.

[link] shows a **relative frequency distribution** of the means based on [link]. This distribution is also a **probability distribution** since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.



Distribution of means for $N = 2$.

The distribution shown in [link] is called the **sampling distribution of the mean**. Specifically, it is the sampling distribution of the mean for a sample size of 2 ($N = 2$). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distribution. The pool balls have only the numbers 1, 2, and 3, and a sample mean can have one of only five possible values.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement) and the mean of the two balls is computed and recorded. Then this process is repeated for a second sample, a third sample, and eventually thousands of a samples. After thousands of samples are taken and the mean computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution shown in [link]. As the number of samples approaches infinity, the frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a frequency distribution based on a very large number of samples. To be strictly correct, the sampling distribution only equals the frequency distribution exactly when there is an infinite number of samples.

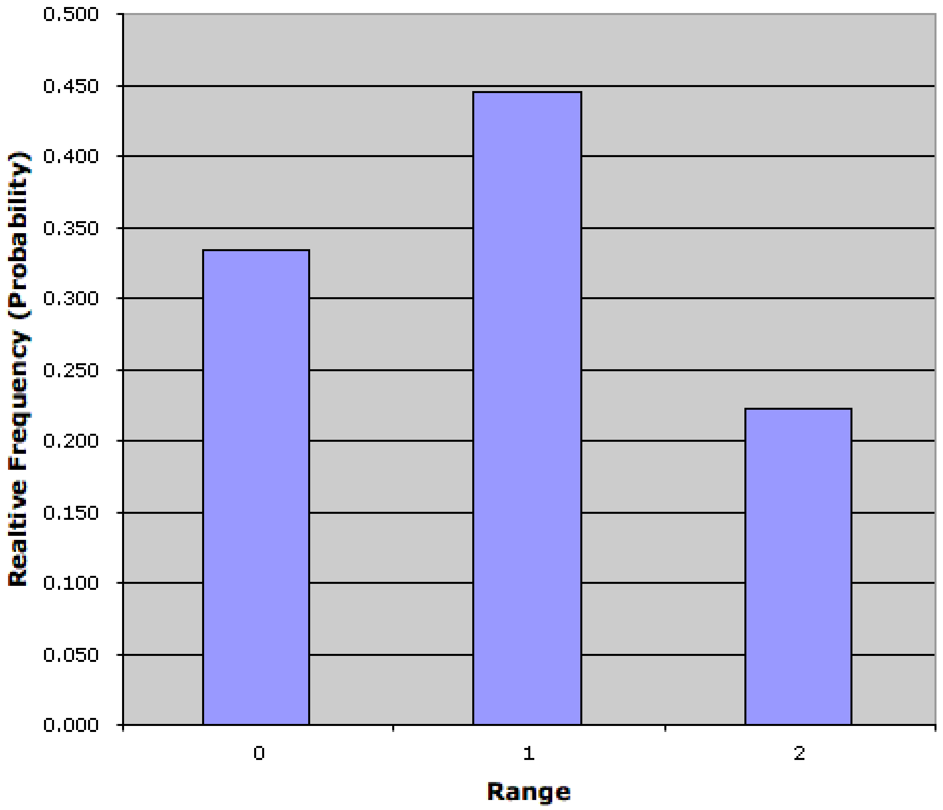
It is important to keep in mind that every statistic, not just the mean has a sampling distribution. For example, [link] shows all possible outcomes for the range of two numbers (larger number minus the smaller number). [link] shows the frequencies for each of the possible ranges and [link] shows the sampling distribution of the range.

Outcome	**Ball 1**	**Ball 2**	**Range**
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

All possible outcomes when two balls are sampled.

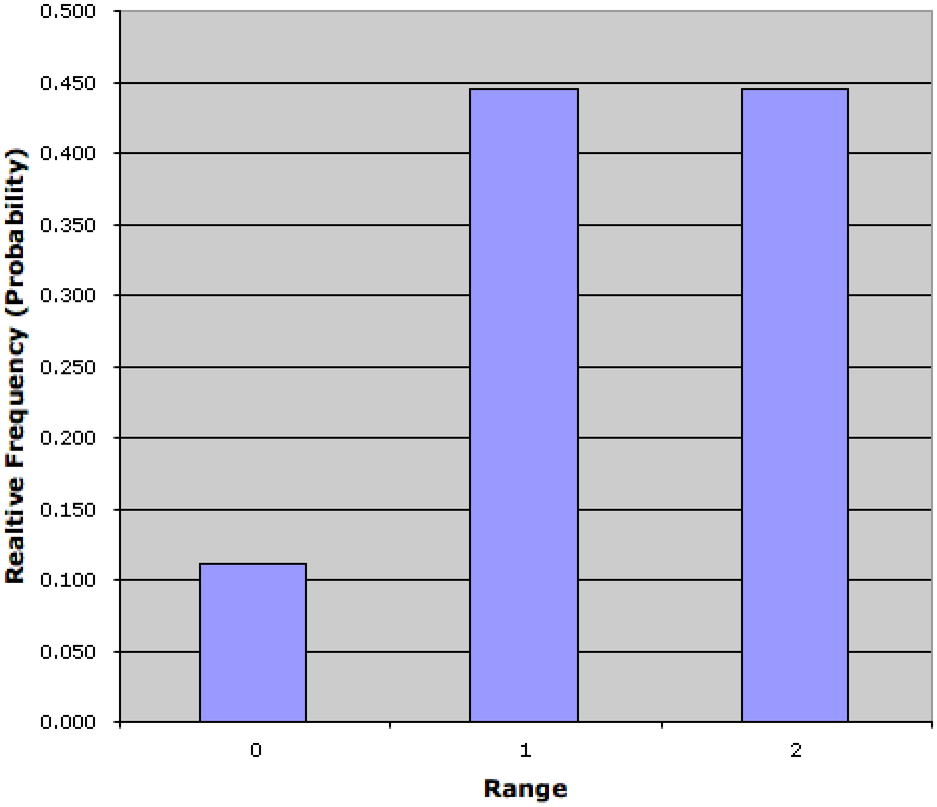
Range	**Frequency**	**Relative Frequency**
0	3	0.333
1	4	0.444
2	2	0.222

Frequencies of ranges for N = 2.



Distribution of ranges for $N = 2$.

It is also important to keep in mind that there is a sampling distribution for various sample sizes. For simplicity, we have been using $N = 2$. The sampling distribution of the range for $N = 3$ is shown in [link].



Distribution of ranges for $N = 3$.

Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps. (Although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this problem since there are 1,000,000 possible outcomes (1,000 for the first ball x 1,000 for the second.) Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions. Specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous distributions, the probability of obtaining any single value is zero. Therefore, as discussed in our introduction to Distributions, these values are called **probability densities** rather than probabilities.

Sampling Distributions and Inferential Statistics

As we stated in the beginning of this chapter, sampling distributions are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: you collect sample data and, from these data, you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution: The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the

standard error of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Keep in mind that all statistics have sampling distributions, not just the mean. In later sections we will be discussing the sampling distribution of the variance, the sampling distribution of the difference between means, and the sampling distribution of Pearson's correlation, among others.

The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

- **a**μ_X = the mean of X
- **b**σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable X which consists of sample means, tends to be **normally distributed** and

$$X \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

The Central Limit Theorem for Sample Means says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n, the sample size. n is the number of values that are averaged together not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n,the distribution of the random variable X, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n, the sample size, increases.

The random variable X has a different z-score associated with it than the random variable X. x is the value of X in one sample.
Equation:

$$z = \frac{x - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$$

μ_X is both the average of X and of X.

$\sigma_X = \frac{\sigma_X}{\sqrt{n}}$ = standard deviation of X and is called the **standard error of the mean.**

Example:
An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population.
Exercise:

Problem:

Find the probability that the **sample mean** is between 85 and 92.

Solution:

Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

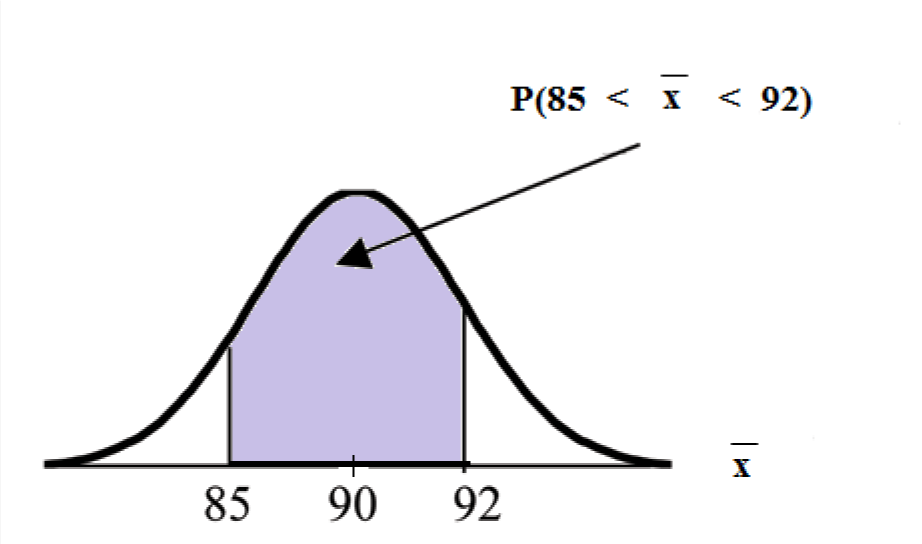
Let X = the mean of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$;

then $X \sim N\left(90, \frac{15}{\sqrt{25}}\right)$

Find $P(85 < x < 92)$ Draw a graph.

$P(85 < x < 92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.



TI-83 or 84: `normalcdf`(lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf`$(85,92,90,\frac{15}{\sqrt{25}}) = 0.6997$

Exercise:

Problem:

Find the value that is 2 standard deviations above the expected value (it is 90) of the sample mean.

Solution:

To find the value that is 2 standard deviations above the expected value 90, use the formula

value = $\mu_X + (\#\text{ofSTDEVs})\left(\frac{\sigma_X}{\sqrt{n}}\right)$

value = $90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$

So, the value that is 2 standard deviations above the expected value is 96.

Example:
The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size** n **= 50** is drawn randomly from the population.
Exercise:

Problem:

Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution:

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let X = the **mean** time, in hours, it takes to play one soccer match.

If $\mu_X =$ ___________, $\sigma_X =$ ____________, and $n =$ _____________, then X~N(_______, _______) by the Central Limit Theorem for Means.

$\mu_X = \mathbf{2}$, $\sigma_X = \mathbf{0.5}$, $n = \mathbf{50}$, and X~$N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < x < 2.3)$.   Draw a graph.

$P(1.8 < x < 2.3) = 0.9977$

`normalcdf`$(1.8{,}2.3{,}2{,}\frac{.5}{\sqrt{50}}) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is _______.

Glossary

Average
: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem
: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(\mathrm{n\mu}, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution
: A continuous random variable (RV) with pdf $\mathrm{f(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Error of the Mean
: The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$.

The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

- **a** $\mu_X =$ the mean of X
- **b** $\sigma_X =$ the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

$$\Sigma X \sim N\left(n \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right)$$

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution) which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

- **a**Σx is one sum.
- **b**$z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$

- **a**$n \cdot \mu_X =$ the mean of ΣX
- **b** $\sqrt{n} \cdot \sigma_X =$ standard deviation of ΣX

Example:
An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.
Exercise:

Problem:

- **a**Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
- **b**Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution:

Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values.**

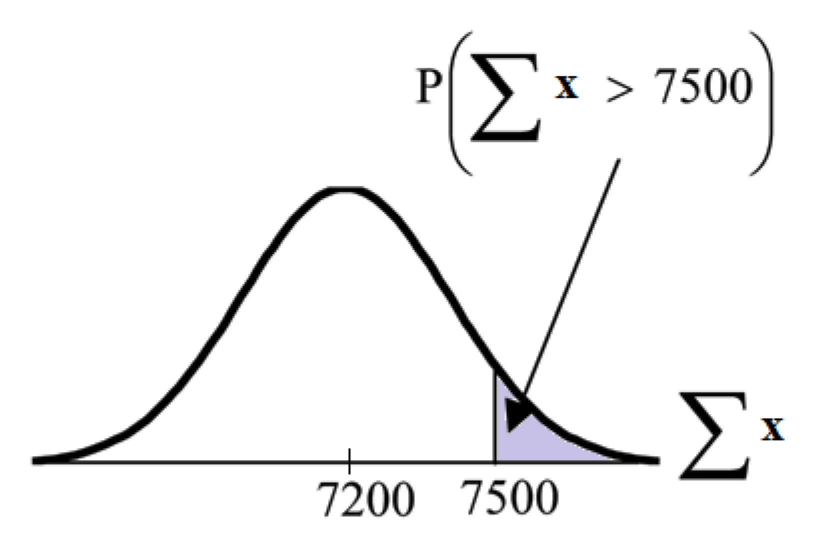
ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then

$$\Sigma X \sim N\left(80 \cdot 90, \sqrt{80} \cdot 15\right)$$

- mean of the sums = $n \cdot \mu_X = (80)(90) = 7200$
- standard deviation of the sums = $\sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
- sum of 80 values = $\Sigma\text{x} = 7500$

- **a**Find $P(\Sigma\text{x} > 7500)$

$P(\Sigma\text{x} > 7500) = 0.0127$



`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated (lower, upper, $n \cdot \mu_X$, $\sqrt{n} \cdot \sigma_X$)

`normalcdf`(7500,1E99, $80 \cdot 90, \sqrt{80} \cdot 15\Big) = 0.0127$

Reminder: 1E99 $= 10^{99}$. Press the `EE` key for E.

- **b**Find Σx where z = 1.5:

Σx = $n \cdot \mu_X + z \cdot \sqrt{n} \cdot \sigma_X$ = (80)(90) + (1.5)($\sqrt{80}$) (15) = 7401.2

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(\text{n}\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution

A continuous random variable (RV) with pdf $\mathrm{f(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: X ~ $N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Using the Central Limit Theorem
Central Limit Theorem: Using the Central Limit Theorem is part of the collection col10555 written by Barbara Illowsky and Susan Dean. It covers how and when to use the Central Limit Theorem and has contributions from Roberta Bloom.

It is important for you to understand when to use the **CLT**. If you are being asked to find the probability of the mean, use the CLT for the mean. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for means and sums.

Note: If you are being asked to find the probability of an **individual** value, do **not** use the CLT. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean x of the sample tends to get closer and closer to μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for X is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean x must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean and Sum Examples

Example:

A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the **mean stress score** for the 75 students is less than 2.
2. The 90th percentile for the **mean stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.
Problems 1. and 2. ask you to find a probability or a percentile for a **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.
Since the individual stress scores follow a uniform distribution, X ~ $U(1, 5)$ where $a = 1$ and $b = 5$ (See Continuous Random Variables for the uniform).

$$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$$

For problems 1. and 2., let X = the mean stress score for the 75 students. Then,

X ~ $N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where n $= 75$.

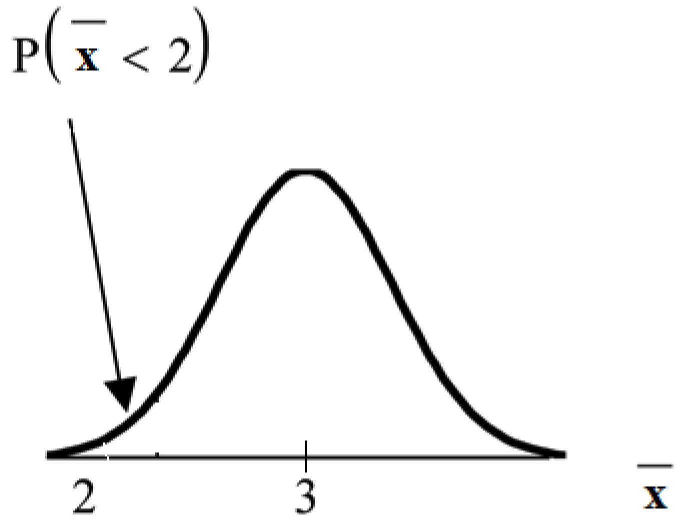
Exercise:

Problem: Find $P(x < 2)$. Draw the graph.

Solution:

$P(x < 2) = 0$

The probability that the mean stress score is less than 2 is about 0.



$\texttt{normalcdf}\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

Note: The smallest stress score is 1. Therefore, the smallest mean for 75 stress scores is 1.

Exercise:

Problem:

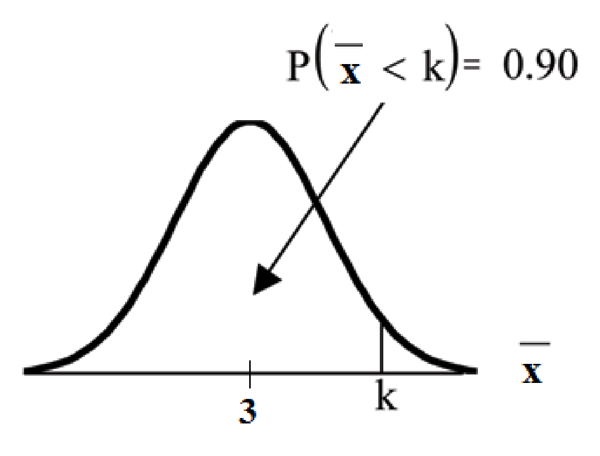
Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

Solution:

Let k = the 90th precentile.

Find k where $P(x < k) = 0.90$.

$k = 3.2$



The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2 and 10% are at least 3.2.

`invNorm` $\left(.90, 3, \frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then, ΣX ~ $N\left[\left(75\right) \cdot \left(3\right), \sqrt{75} \cdot 1.15\right]$

Exercise:

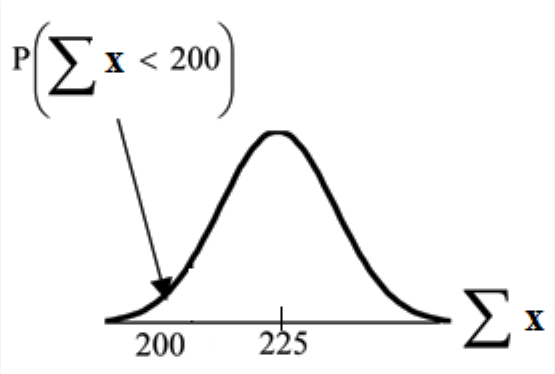
Problem: Find $P(\Sigma\text{x} < 200)$. Draw the graph.

Solution:

The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75} \cdot 1.15 = 9.96$

$P(\Sigma\text{x} < 200) = 0$



The probability that the total of 75 scores is less than 200 is about 0.

normalcdf $\left(75, 200, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 0.$

Note: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Exercise:

Problem:

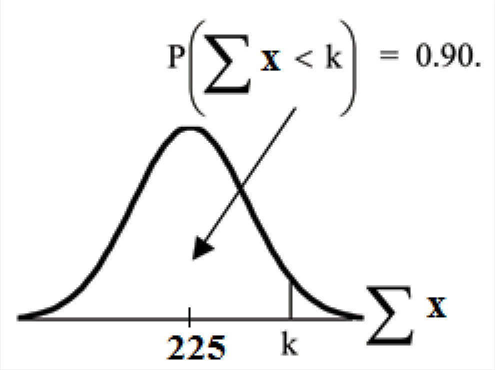
Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution:

Let k = the 90th percentile.

Find k where $P(\Sigma\text{x} < k) = 0.90.$

$k = 237.8$



The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm` $\left(.90, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 237.8$

Example:

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim \text{Exp}\left(\frac{1}{22}\right)$ From Chapter 5, we know that $\mu = 22$ and $\sigma = 22$.

Let X = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$X \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means

Exercise:

Problem:

Using the CLT to find Probability:

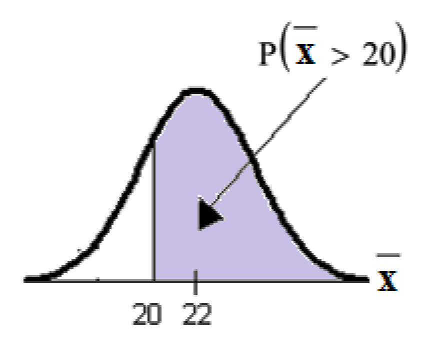
- **a**Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(x > 20)$ Draw the graph.
- **b** Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$
- **c** Explain why the probabilities in (a) and (b) are different.

Solution:

Part a.
Find: $P(x > 20)$

$P(x > 20) = 0.7919$ using `normalcdf` $\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



Note:$1E99 = 10^{99}$and-$1E99 = -10^{99}$. Press the

`EE`

key for E. Or just use 10^99 instead of 1E99.

Part b.
Find P(x>20) . Remember to use the exponential distribution for an **individual: X~Exp(1/22)**.

P(X>20) = e^(–(1/22)*20) or e^(–.04545*20) = 0.4029

Part c. Explain why the probabilities in (a) and (b) are different.

- $P(x > 20) = 0.4029$ but $P(x > 20) = 0.7919$
- The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
- When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the CLT. Use the CLT with the normal distribution when you are being asked to find the probability for an mean.

Exercise:

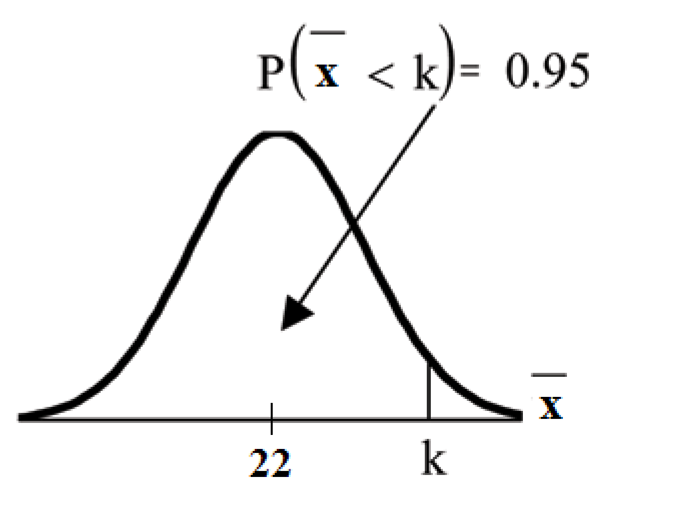
Problem:

Using the CLT to find Percentiles:
Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution:

Let k = the 95th percentile. Find k where $P(x < k) = 0.95$

$k = 26.0$ using $\texttt{invNorm}\left(.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$



The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

95% of such samples would have means under 26 minutes; only 5% of such samples would have means above 26 minutes.

Note:(HISTORICAL): Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for n (say, 20). To calculate the probabilities with large values of n, you had to use the binomial formula which could be very complicated. Using the **Normal Approximation to the Binomial** simplified the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure

- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then X~$B(n,p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x $($use $x + 0.5$ or $x - 0.5$ $)$. The number 0.5 is called the **continuity correction factor**.

Example:
Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K - 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let X= the number that favor a charter school for grades K - 5. X~ $B(n,p)$ where n =300 and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. Y~$N(159, 8.6447)$. See **The Normal Distribution** for help with calculator instructions.
For Problem 1., you **include 150** so $P(x \geq 150)$ has normal approximation $P(Y \geq 149.5)$=0.8641.
`normalcdf` $(149.5, 10\hat{}99, 159, 8.6447) = 0.8641$.
For Problem 2., you **include 160** so $P(x \leq 160)$ has normal approximation $P(Y \leq 160.5)$=0.5689.

`normalcdf` $(0,160.5,159,8.6447)=0.5689$
For Problem 3., you **exclude 155** so $P(x>155)$ has normal approximation $P(y>155.5)$=0.6572.
`normalcdf` $(155.5,10\text{ˆ}99,159,8.6447)=0.6572$
For Problem 4., you **exclude 147** so $P(x<147)$ has normal approximation $P(Y<146.5)$=0.0741.
`normalcdf` $(0,146.5,159,8.6447)=0.0741$
For Problem 5., $P(x$=175$)$ has normal approximation $P(174.5<y<175.5)$=0.0083.
`normalcdf` $(174.5,175.5,159,8.6447)=0.0083$
Because of calculators and computer software that easily let you calculate binomial probabilities for large values of n, it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial. In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.
For **Example 3**, the probabilities are calculated using the binomial (n=300 and p=0.53) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.
$P(x\geq 150)$: `1 - binomialcdf` $(300,0.53,149)$=0.8641
$P(x\leq 160)$: `binomialcdf` $(300,0.53,160)$=0.5684
$P(x>155)$: `1 - binomialcdf` $(300,0.53,155)$=0.6576
$P(x<147)$: `binomialcdf` $(300,0.53,146)$=0.0742
$P(x$=175$)$: (You use the binomial pdf.) `binomialpdf` $(175,0.53,146)$=0.0083

**Contributions made to Example 2 by Roberta Bloom

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\mathit{\Sigma}X$. If the size n of the sample is sufficiently large, then $\bar{X}\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\mathit{\Sigma}X \sim N\left(\text{nμ}, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Exponential Distribution

A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: X~$\text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = \text{me}^{-\text{mx}}, x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-\text{mx}}$.

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Uniform Distribution

A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X\text{~}U(a{,}b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \le x \le b$. The cumulative distribution is $P(X \le x) = \frac{x-a}{b-a}$.

Summary of Formulas

Formula

Central Limit Theorem for Sample Means

$X \sim \mathrm{N}\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ **The Mean** $\left(X\right)$**:** μ_X

Formula

Central Limit Theorem for Sample Means Z-Score and Standard Error of the Mean

$z = \frac{x-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$ **Standard Error of the Mean (Standard Deviation** $\left(X\right)$**):** $\frac{\sigma_X}{\sqrt{n}}$

Formula

Central Limit Theorem for Sums

$\Sigma\mathrm{X} \sim N\left[\left(n\right) \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right]$ **Mean for Sums** $(\Sigma\mathrm{X})$**:** $n \cdot \mu_X$

Formula

Central Limit Theorem for Sums Z-Score and Standard Deviation for Sums

$z = \frac{\Sigma\mathrm{x}-n\cdot\mu_X}{\sqrt{n}\cdot\sigma_X}$ **Standard Deviation for Sums** $(\Sigma\mathrm{X})$**:** $\sqrt{n} \cdot \sigma_X$

Practice: The Central Limit Theorem

Student Learning Outcomes

- The student will calculate probabilities using the Central Limit Theorem.

Given

Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately 4 hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let X be the random variable representing the mean time to complete the 16 reviews. Let ΣX be the total time it takes Yoonie to complete all of the month’s reviews. Assume that the 16 reviews represent a random set of reviews.

Distribution

Complete the distributions.

1. X ~
2. X ~
3. ΣX ~

Graphing Probability

For each problem below:

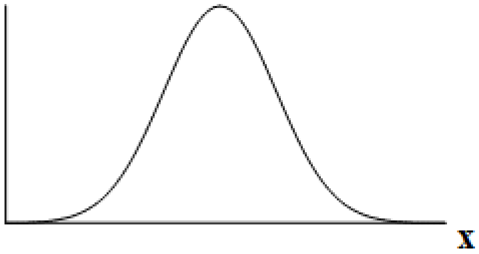
- **a** Sketch the graph. Label and scale the horizontal axis. Shade the region corresponding to the probability.
- **b** Calculate the value.

Exercise:

Problem:

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.

- **a**



- **b** $P($ ________ $< x <$ ________ $) =$ _______

Solution:

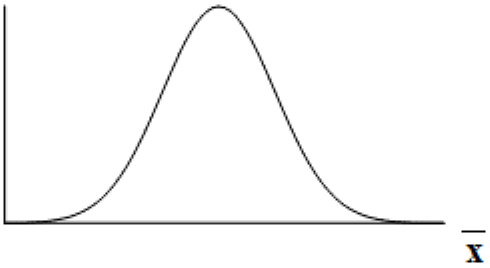
- **b** 3.5, 4.25, 0.2441

Exercise:

Problem:

Find the probability that the **mean** of a month’s reviews will take Yoonie from 3.5 to 4.25 hrs.

- **a**



- **b** $P(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_) = \_\_\_\_\_\_\_$

Solution:

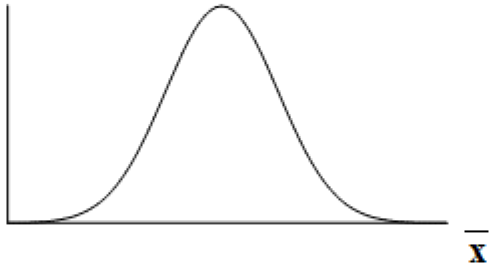
- **b** 0.7499

Exercise:

Problem:

Find the 95th percentile for the **mean** time to complete one month's reviews.

- **a**



- **b**The 95th Percentile=

Solution:

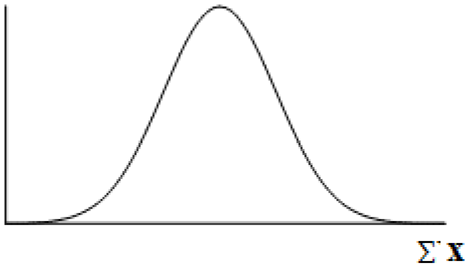
- **b** 4.49 hours

Exercise:

Problem:

Find the probability that the **sum** of the month's reviews takes Yoonie from 60 to 65 hours.

- **a**



- **b** The Probability=

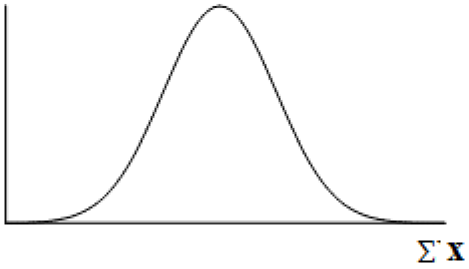
Solution:

- **b** 0.3802

Exercise:

Problem: Find the 95th percentile for the **sum** of the month’s reviews.

- **a**



- **b** The 95th percentile=

Solution:

- **b** 71.90

Discussion Question

Exercise:

Problem: What causes the probabilities in [link] and [link] to differ?

Homework

The Central Limit Theorem: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean.

Exercise:

Problem:

X ~ $N(60{,}9)$. Suppose that you form random samples of 25 from this distribution. Let X be the random variable of averages. Let ΣX be the random variable of sums. For **c - f**, sketch the graph, shade the region, label and scale the horizontal axis for X, and find the probability.

- **a** Sketch the distributions of X and X on the same graph.
- **b** X ~
- **c** $P(x < 60) =$
- **d** Find the 30th percentile for the mean.
- **e** $P(56 < x < 62) =$
- **f** $P(18 < x < 58) =$
- **g** Σx ~
- **h** Find the minimum value for the upper quartile for the sum.
- **i** $P(1400 < \Sigma x < 1550) =$

Solution:

- **b** Xbar~$N(60, \frac{9}{\sqrt{25}})$
- **c**0.5000
- **d**59.06
- **e**0.8536
- **f**0.1333
- **h**1530.35
- **i**0.8536

Exercise:

Problem:

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

- **a** When the sample size is large, the mean of X is approximately equal to the mean of X.
- **b** When the sample size is large, X is approximately normally distributed.
- **c** When the sample size is large, the standard deviation of X is approximately the same as the standard deviation of X.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about 10. Suppose that 16 individuals are randomly chosen.

Let X =average percent of fat calories.

- **a** X ~______ (______ , ______)
- **b** For the group of 16, find the probability that the average percent of fat calories consumed is more than 5. Graph the situation and shade in the area to be determined.
- **c** Find the first quartile for the average percent of fat calories.

Solution:

- **a** $N(36, \frac{10}{\sqrt{16}})$
- **b** 1
- **c** 34.31

Exercise:

Problem:

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of $0.88. Suppose that we randomly pick 25 daytime statistics students.

- **a** In words, $X =$
- **b** X~
- **c** In words, $X =$
- **d** X~ ______ (______ , ______)
- **e** Find the probability that an individual had between $0.80 and $1.00. Graph the situation and shade in the area to be determined.
- **f** Find the probability that the average of the 25 students was between $0.80 and $1.00. Graph the situation and shade in the area to be determined.
- **g** Explain the why there is a difference in (e) and (f).

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

- **a** If $X =$ average distance in feet for 49 fly balls, then X~ _______ (_______ , _______)
- **b** What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for X. Shade the region corresponding to the probability. Find the probability.
- **c** Find the 80th percentile of the distribution of the average of 49 fly balls.

Solution:

- **a** $N(250, \frac{50}{\sqrt{49}})$
- **b** 0.0808
- **c**256.01 feet

Exercise:

Problem:

Suppose that the weight of open boxes of cereal in a home with children is uniformly distributed from 2 to 6 pounds. We randomly survey 64 homes with children.

- **a** In words, $X =$
- **b** X~
- **c** $\mu_X =$
- **d** $\sigma_X =$
- **e** In words, $\Sigma X =$
- **f** ΣX~
- **g** Find the probability that the total weight of open boxes is less than 250 pounds.
- **h** Find the 35th percentile for the total weight of open boxes of cereal.

Exercise:

Problem:

Suppose that the duration of a particular type of criminal trial is known to have a mean of 21 days and a standard deviation of 7 days. We randomly sample 9 trials.

- **a** In words, $\Sigma X =$
- **b** ΣX~
- **c** Find the probability that the total length of the 9 trials is at least 225 days.
- **d** 90 percent of the total of 9 of these types of trials will last at least how long?

Solution:

- **a**The total length of time for 9 criminal trials
- **b** $N(189{,}21)$
- **c** 0.0432
- **d** 162.09

Exercise:

Problem:

According to the Internal Revenue Service, the average length of time for an individual to complete (record keep, learn, prepare, copy, assemble and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is 2 hours. Suppose we randomly sample 36 taxpayers.

- **a** In words, $X =$
- **b** In words, $X =$
- **c** X~
- **d** Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
- **e** Would you be surprised if one taxpayer finished his Form 1040 in more than 12 hours? In a complete sentence, explain why.

Exercise:

Problem:

Suppose that a category of world class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races.

Let $X =$ the average of the 49 races.

- **a** X~
- **b** Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
- **c** Find the 80th percentile for the average of these 49 marathons.
- **d** Find the median of the average running times.

Solution:

- **a** $N(145, \frac{14}{\sqrt{49}})$
- **b** 0.6247
- **c** 146.68
- **d** 145 minutes

Exercise:

Problem:

The attention span of a two year-old is exponentially distributed with a mean of about 8 minutes. Suppose we randomly survey 60 two year-olds.

- **a** In words, $X =$
- **b** X~
- **c** In words, $X =$
- **d** X~
- **e** Before doing any calculations, which do you think will be higher? Explain why.
  - **i** the probability that an individual attention span is less than 10 minutes; or
  - **ii** the probability that the average attention span for the 60 children is less than 10 minutes? Why?
- **f** Calculate the probabilities in part (e).
- **g** Explain why the distribution for X is not exponential.

Exercise:

Problem:

Suppose that the length of research papers is uniformly distributed from 10 to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.

- **a** In words, $X =$
- **b** X~
- **c** $\mu_X =$
- **d** $\sigma_X =$
- **e** In words, $X =$
- **f** X~
- **g** In words, $\Sigma X =$
- **h** ΣX~
- **i** Without doing any calculations, do you think that it's likely that the professor will need to read a total of more than 1050 pages? Why?
- **j** Calculate the probability that the professor will need to read a total of more than 1050 pages.
- **k** Why is it so unlikely that the average length of the papers will be less than 12 pages?

Solution:

- **b** $U(10{,}25)$
- **c**17.5
- **d** $\sqrt{\frac{225}{12}}$ = 4.3301
- **f**$N(17.5{,}0.5839)$
- **h**$N(962.5{,}32.11)$
- **j**0.0032

Exercise:

Problem:

The length of songs in a collector's CD collection is uniformly distributed from 2 to 3.5 minutes. Suppose we randomly pick 5 CDs from the collection. There is a total of 43 songs on the 5 CDs.

- **a** In words, $X =$
- **b** X~
- **c** In words, X=
- **d** X~
- **e** Find the first quartile for the average song length.
- **f** The IQR (interquartile range) for the average song length is from _______ to _______.

Exercise:

Problem:

Salaries for teachers in a particular elementary school district are normally distributed with a mean of $44,000 and a standard deviation of $6500. We randomly survey 10 teachers from that district.

- **a** In words, $X =$
- **b** In words, $X =$
- **c** X~
- **d** In words, $\Sigma X =$
- **e** ΣX~
- **f** Find the probability that the teachers earn a total of over $400,000.
- **g** Find the 90th percentile for an individual teacher's salary.
- **h** Find the 90th percentile for the average teachers' salary.
- **i** If we surveyed 70 teachers instead of 10, graphically, how would that change the distribution for X?
- **j** If each of the 70 teachers received a $3000 raise, graphically, how would that change the distribution for X?

Solution:

- **c** $N(44{,}000, \frac{6500}{\sqrt{10}})$
- **e** $N(440{,}000, (\sqrt{10})(6500))$
- **f** 0.9742
- **g** $52,330
- **h** $46,634

Exercise:

Problem:

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and few to many wealthy people). Suppose we pick a country with a wedge distribution. Let the average salary be $2000 per year with a standard deviation of $8000. We randomly survey 1000 residents of that country.

- **a** In words, $X =$
- **b** In words, $X =$
- **c** X ~
- **d** How is it possible for the standard deviation to be greater than the average?
- **e** Why is it more likely that the average of the 1000 residents will be from $2000 to $2100 than from $2100 to $2200?

Exercise:

Problem:

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

- **a** In words, $X =$
- **b** In words, $X =$

- **c** X~
- **d** In words, $\Sigma X =$
- **e** ΣX~
- **f** Is it likely that an individual stayed more than 5 days in the hospital? Why or why not?
- **g** Is it likely that the average stay for the 80 women was more than 5 days? Why or why not?
- **h** Which is more likely:
    - **i**an individual stayed more than 5 days; or
    - **ii**the average stay of 80 women was more than 5 days?
- **i** If we were to sum up the women's stays, is it likely that, collectively they spent more than a year in the hospital? Why or why not?

Solution:

- **c** $N(2.4, \frac{0.9}{\sqrt{80}})$
- **e**$N(192{,}8.05)$
- **h**Individual

Exercise:

Problem:

In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940. (Source: U.S. Dept. of Agriculture)

- **a** In words, $X =$
- **b** In words, $X =$
- **c** X~
- **d** The IQR for X is from _______ acres to _______ acres.

Exercise:

Problem:

The stock closing prices of 35 U.S. semiconductor manufacturers are given below. (Source: **Wall Street Journal**)

8.625 30.25 27.625 46.75 32.875 18.25 5 0.125 2.9375 6.875 28.25 24.25 21 1.5 30.25 71 43.5 49.25 2.5625 31 16.5 9.5 18.5 18 9 10.5 16.625 1.25 18 12.875 7 12.875 2.875 60.25 29.25

- **a** In words, $X =$
- **b**
    - **i**$x =$
    - **ii** $s_x =$
    - **iii** $n =$
- **c** Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
- **d** In words, describe the distribution of stock prices.
- **e** Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 10 averages. List those 10 averages.
- **f** Use the 10 averages from (e) to calculate:
    - **i** $x =$
    - **ii** $s_x =$
- **g** Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
- **h** Does this histogram look like the graph in (c)?
- **i** In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
- **j** Based upon the theory of the Central Limit Theorem, X~

Solution:

- **b**$20.71; $17.31; 35

- **d** Exponential distribution, X ~ $\text{Exp}(1/20.71)$
- **f** \$20.71; \$11.14
- **j** $N(20.71, \frac{17.31}{\sqrt{5}})$

Exercise:

Problem:

Use the Initial Public Offering data (see “Table of Contents) to do this problem.

- **a** In words, $X =$
- **b**
  - **i**$\mu_X =$
  - **ii** $\sigma_X =$
  - **iii** $n =$
- **c** Construct a histogram of the distribution. Start at $x = -0.50$. Make bar widths of \$5.
- **d** In words, describe the distribution of stock prices.
- **e** Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 15 averages. List those 15 averages.
- **f** Use the 15 averages from (e) to calculate the following:
  - **i**$x =$
  - **ii** $s_x =$
- **g** Construct a histogram of the distribution of the averages. Start at $x = -0.50$. Make bar widths of \$5.
- **h** Does this histogram look like the graph in (c)? Explain any differences.
- **i** In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
- **j** Based upon the theory of the Central Limit Theorem, X~

Try these multiple choice questions (Exercises19 - 23).

The next two questions refer to the following information: The time to wait for a particular rural bus is distributed uniformly from 0 to 75 minutes. 100 riders are randomly sampled to learn how long they waited.

Exercise:

Problem:

The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

- **A** 315.0
- **B** 40.3
- **C** 38.5
- **D** 65.2

Solution:

B

Exercise:

Problem:

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

- **A** Yes
- **B** No
- **C** There is not enough information.

Solution:

A

Exercise:

Problem:

Which of the following is NOT TRUE about the distribution for averages?

- **A** The mean, median and mode are equal
- **B** The area under the curve is one
- **C** The curve never touches the x-axis
- **D** The curve is skewed to the right

Solution:

D

The next three questions refer to the following information: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

Exercise:

Problem:

The distribution to use for the average cost of gasoline for the 16 gas stations is

- **A** $X \sim N(4.59, 0.10)$
- **B** $X \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
- **C** $X \sim N\left(4.59, \frac{0.10}{16}\right)$
- **D** $X \sim N\left(4.59, \frac{16}{0.10}\right)$

Solution:

B

Exercise:

Problem:

What is the probability that the average price for 16 gas stations is over $4.69?

- **A** Almost zero
- **B** 0.1587
- **C** 0.0943
- **D** Unknown

Solution:

A

Exercise:

Problem:

Find the probability that the average price for 30 gas stations is less than $4.55.

- **A**0.6554
- **B**0.3446
- **C**0.0142
- **D**0.9858
- **E**0

Solution:

C

Exercise:

Problem:

For the Charter School Problem (Example 6) in **Central Limit Theorem: Using the Central Limit Theorem**, calculate the following using the normal approximation to the binomial.

- **A** Find the probability that less than 100 favor a charter school for grades K - 5.
- **B** Find the probability that 170 or more favor a charter school for grades K - 5.
- **C** Find the probability that no more than 140 favor a charter school for grades K - 5.
- **D** Find the probability that there are fewer than 130 that favor a charter school for grades K - 5.
- **E** Find the probability that exactly 150 favor a charter school for grades K - 5.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Solution:

- **C** 0.0162
- **E** 0.0268

Exercise:

Problem:

Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to 4 decimal places.

- **A** Find the probability that Janice is the driver at most 20 days.
- **B** Find the probability that Roberta is the driver more than 16 days.
- **C** Find the probability that Barbara drives exactly 24 of those 96 days.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Solution:

- **A** 0.2047
- **B** 0.9615
- **C** 0.0938

**Exercise 24 contributed by Roberta Bloom

Review
Central Limit Theorem: Review is part of the collection col10555 written by Barbara Illowsky and Susan Dean. The module consists of review exercises.

The next three questions refer to the following information: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

Exercise:

Problem: X ~

- **A** $U(0,4)$
- **B** $U(10,2)$
- **C** $\mathrm{Exp}(2)$
- **D** $N(2,1)$

Solution:

A

Exercise:

Problem: The average wait time is:

- **A** 1 hour
- **B** 2 hour
- **C** 2.5 hour
- **D** 4 hour

Solution:

B

Exercise:

Problem:

Suppose that it is now past noon on a delivery day. The probability that a person must wait at least $1\frac{1}{2}$ **more** hours is:

- **A** $\frac{1}{4}$
- **B** $\frac{1}{2}$
- **C** $\frac{3}{4}$
- **D** $\frac{3}{8}$

Solution:

A

Exercise:

Problem: Given: X~Exp$\left(\frac{1}{3}\right)$.

- **a** Find $P(x > 1)$
- **b** Calculate the minimum value for the upper quartile.
- **c** Find $P\left(x = \frac{1}{3}\right)$

Solution:

- **a** 0.7165
- **b** 4.16
- **c** 0

Exercise:

Problem:

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate

- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

- **A** 4 years
- **B** 4.5 years
- **C** 5 years
- **D** 5.5 years

Solution:

C

Exercise:

Problem:

Which of the following distributions is described by the following example?

Many people can run a short distance of under 2 miles, but as the distance increases, fewer people can run that far.

- **A** Binomial
- **B** Uniform
- **C** Exponential
- **D** Normal

Solution:

C

Exercise:

Problem:

The length of time to brush one’s teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his/her teeth less than $\frac{3}{4}$ minutes.

- **A** 0.5
- **B** $\frac{3}{4}$
- **C** 0.43
- **D** 0.63

Solution:

D

Exercise:

Problem:

Which distribution accurately describes the following situation?

The chance that a teenage boy regularly gives his mother a kiss goodnight (and he should!!) is about 20%. Fourteen teenage boys are randomly surveyed.

$X =$the number of teenage boys that regularly give their mother a kiss goodnight

- **A** $B(14,0.20)$
- **B** $P(2.8)$
- **C** $N(2.8,2.24)$
- **D** $\text{Exp}(\frac{1}{0.20})$

Solution:

A

Exercise:

Problem:

Which distribution accurately describes the following situation?

A 2008 report on technology use states that approximately 20 percent of U.S. households have never sent an e-mail. (source: http://www.webguild.org/2008/05/20-percent-of-americans-have-never-used-email.php) Suppose that we select a random sample of fourteen U.S. households .

$X =$the number of households in a 2008 sample of 14 households that have never sent an email

- **A** $B(14{,}0.20)$
- **B** $P(2.8)$
- **C** $N(2.8{,}2.24)$
- **D** $\text{Exp}(\frac{1}{0.20})$

Solution:

A

**Exercise 9 contributed by Roberta Bloom

Introduction

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for one population mean and one population proportion.
- Interpret the student-t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the student-t distributions.

Introduction

Suppose you are trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percent of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct confidence intervals in which we believe the parameter lies.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of compact discs (CD's) a consumer buys per month. If so, you could conduct a survey and calculate the sample mean, $\overline{x}$, and the sample standard deviation, s. You would use $\overline{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\overline{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\overline{x}$ and s is also called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the CD example we do not know the population mean μ but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then by the Central Limit Theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, will be within two standard deviations of the population mean μ. For our CD example, two standard deviations is $(2)(0.1) = 0.2$. The sample mean $\overline{x}$ is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $((2)(0.1))$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\overline{x} - 0.2$ and $\overline{x} + 0.2$ in 95% of all the samples.

For the CD example, suppose that a sample produced a sample mean $\overline{x} = 2$. Then the unknown population mean μ is between

$\overline{x} - 0.2 = 2 - 0.2 = 1.8$ and $\overline{x} + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the unknown population mean number of CDs is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (100% - 95%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form

(point estimate - margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate + or - the margin of error. These are two ways of expressing the same concept.

Note: Although the text only covers symmetric confidence intervals, there are non-symmetric confidence intervals (for example, a confidence interval for the standard deviation).

Optional Collaborative Classroom Activity

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be 3 meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. $\sigma = 3$ and $n =$ the number of students surveyed.
3. Construct the interval $\left(\overline{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}\right)$

We say we are approximately 95% confident that the true average number of meals that students eat out in a week is between __________ and ___________.

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Inferential Statistics

Also called statistical inference or inductive statistics. This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if 4 out of the 100 calculators sampled are defective we might infer that 4 percent of the production is defective.

Parameter

A numerical characteristic of the population.

Point Estimate

A single number computed from a sample and used to estimate a population parameter.

Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal
Confidence Intervals: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ **, where the population standard deviation is known,** we need x as an estimate for μ and we need the margin of error. Here, the margin of error is called the **error bound for a population mean** (abbreviated **EBM**). The sample mean x is the **point estimate** of the unknown population mean μ
The confidence interval estimate will have the form:

- (point estimate - error bound, point estimate + error bound) or, in symbols,$(x - \text{EBM}, x + \text{EBM})$

The margin of error depends on the **confidence level** (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level CL. α is the probability that the interval does not contain the unknown population parameter.
Mathematically, α + CL = 1.

Example:

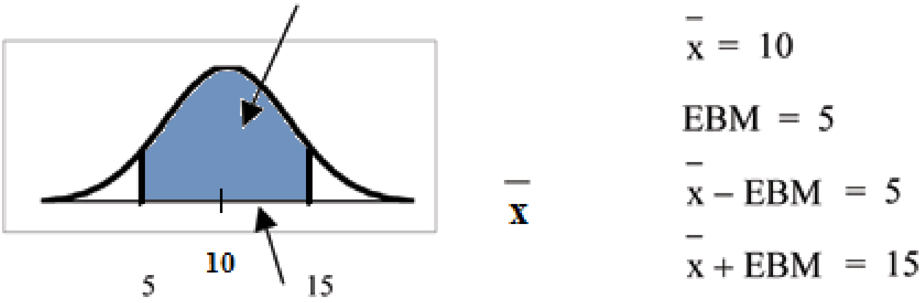
- Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.
- The sample mean is 7 and the error bound for the mean is 2.5.

$x = 7$ and $\text{EBM} = 2.5$.
The confidence interval is $(7 - 2.5, 7 + 2.5)$; calculating the values gives $(4.5, 9.5)$.
If the confidence level (CL) is 95%, then we say that "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $x = 10$ and we have constructed the 90% confidence interval (5, 15) where $\text{EBM} = 5$.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.

Confidence Level (CL) = 0.90



μ is believed to be in the interval (5, 15) with 90% confidence.

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. 1.645 is the z-score from a

Standard Normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating. So in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. $\frac{\sigma}{\sqrt{n}}$ is commonly called the "standard error of the mean" in order to clearly distinguish the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the Central Limit Theorem:

- X is normally distributed, that is, $X \sim \mathrm{N}\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a Normal distribution to calculate the error bound.**

Calculating the Confidence Interval:

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean x from the sample data. Remember, in this section, we already know the population standard deviation σ.
- Find the Z-score that corresponds to the confidence level.
- Calculate the error bound EBM
- Construct the confidence interval
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding z for the stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution Z~N(0,1).

The confidence level, CL, is the area in the middle of the standard normal distribution. $\text{CL} = 1 - \alpha$. So α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$

For example, when $\text{CL} = 0.95$ then $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{.025}$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 1-0.025 = 0.975

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a Standard Normal probability table.

Using the TI83, TI83+ or TI84+ calculator: `invNorm`$(0.975, 0, 1) = 1.96$

CALCULATOR NOTE: Remember to use area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0,1 because you are using a Standard Normal Distribution Z~N(0,1)

EBM: Error Bound
The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

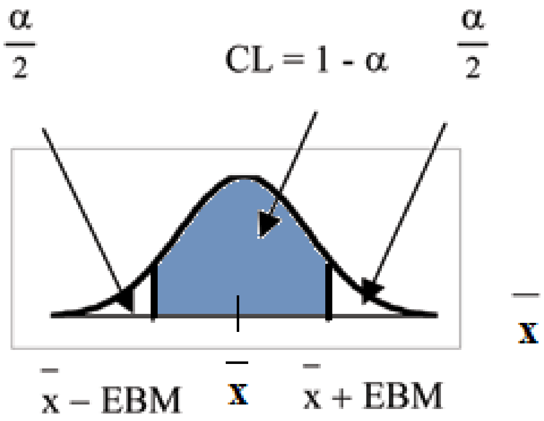
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(x - \text{EBM}, x + \text{EBM})$.

The graph gives a picture of the entire situation.

$\mathrm{CL} + \frac{\alpha}{2} + \frac{\alpha}{2} = \mathrm{CL} + \alpha = 1.$



Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and should state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include context of the problem) is between ___ and ___ (include appropriate units)."

Example:

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Exercise:

Problem:

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution:

- You can use technology to directly calculate the confidence interval
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+ and 84+ calculators (Solution B).

Solution A
To find the confidence interval, you need the sample mean, x, and the EBM.

- $x = 68$
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; $n = 36$; The confidence level is 90% (CL=0.90)

$\text{CL} = 0.90$ so $\alpha = 1 - \text{CL} = 1 - 0.90 = 0.10$

$\frac{\alpha}{2} = 0.05 \qquad z_{\frac{\alpha}{2}} = z_{.05}$

The area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 1−0.05=0.95

$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

using invNorm(0.95,0,1) on the TI-83,83+,84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.

$\text{EBM} = 1.645 \cdot \left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$x - \text{EBM} = 68 - 0.8225 = 67.1775$

$x + \text{EBM} = 68 + 0.8225 = 68.8225$

The 90% confidence interval is **(67.1775, 68.8225).**

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `7:ZInterval`.
Press `ENTER`.
Arrow to `Stats` and press `ENTER`.
Arrow down and enter 3 for σ, 68 for x , 36 for n, and .90 for `C-level`.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (to 3 decimal places) (67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Changing the Confidence Level or Sample Size

Example: Changing the Confidence Level
Exercise:

Problem:

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution:

To find the confidence interval, you need the sample mean, x, and the EBM.

- $x = 68$
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; $n = 36$; The confidence level is 95% (CL=0.95)

$\text{CL} = 0.95$ so $\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \qquad z_{\frac{\alpha}{2}} = z_{.025}$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 1−0.025=0.975

$z_{\frac{\alpha}{2}} = z_{.025} = 1.96$

using invnorm(.975,0,1) on the TI-83,83+,84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.)

$\text{EBM} = 1.96 \cdot \left(\frac{3}{\sqrt{36}}\right) = 0.98$

$x - \text{EBM} = 68 - 0.98 = 67.02$

$x + \text{EBM} = 68 + 0.98 = 68.98$

Interpretation

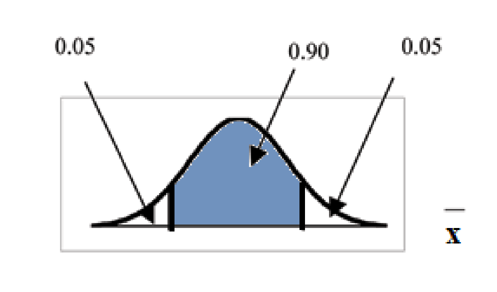
We estimate with 95 % confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

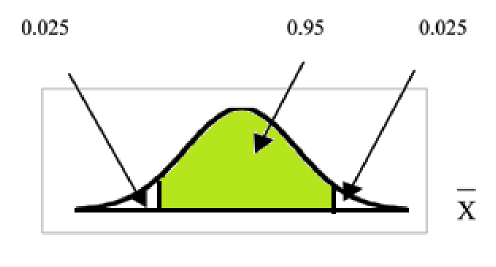
Explanation of 95% Confidence Level

95% of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider.





Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Example: Changing the Sample Size:
Suppose we change the original problem to see what happens to the error bound if the sample size is changed.
Exercise:

Problem:

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use n=100 instead of n=36? What happens if we decrease the sample size to n=25 instead of n=36?

- $x = 68$
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90) ; $z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

Solution:

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100$: $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) = 1.645 \cdot \left(\frac{3}{\sqrt{100}}\right) = 0.4935$

Solution:

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) = 1.645 \cdot \left(\frac{3}{\sqrt{25}}\right) = 0.987$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Working Backwards to Find the Error Bound or Sample Mean

Working Bacwards to find the Error Bound or the Sample Mean

When we calculate a confidence interval, we find the sample mean and calculate the error bound and use them to calculate the confidence interval. But sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean
- OR, From the upper value for the interval, subtract the lower value. Then divide the difference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval
- OR, Average the upper and lower endpoints of the confidence interval

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example:
Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68. Or perhaps our source only gave the confidence interval and did not tell us the value of the the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $\text{EBM} = 68.82 - 68 = 0.82$
- If we don't know the sample mean: $\text{EBM} = \frac{(68.82-67.18)}{2} = 0.82$

Calculate the Sample Mean:

- If we know the error bound: $x = 68.82 - 0.82 = 68$
- If we don't know the error bound: $x = \frac{(67.18+68.82)}{2} = 68$

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$

The formula for sample size is $n = \frac{z^2\sigma^2}{\text{EBM}^2}$, found by solving the error bound formula for n

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example:
The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within 2 years of the true population mean age of Foothill College students , how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and EBM=2
- $z = z_{.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{\text{EBM}^2} = \frac{1.96^2 15^2}{2^2}$ =216.09 using the sample size equation.
- Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within 2 years of the true population mean age of Foothill College students.

**With contributions from Roberta Bloom

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.

- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $\text{CL} = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's-t

Confidence Interval, Single Population Mean, Population Standard Deviation Unknown, Student-t is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gossett (1876-1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's-t distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid 1970s, some statisticians used the **normal distribution** approximation for large sample sizes and only used the Student's-t distribution for sample sizes of at most 30. With the common use of graphing calculators and computers, the practice is to use the Student's-t distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score $\mathrm{t} = \frac{\mathrm{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student's-t distribution with $n-1$ degrees of freedom**. The t-score has the same interpretation as the **z-score**. It measures how far x is from its mean μ. For each sample size n, there is a different Student's-t distribution.

The **degrees of freedom**, $n-1$, come from the calculation of the sample standard deviation s. In Chapter 2, we used n deviations $(x-x\ \text{ values})$ to calculate s. Because the sum of the deviations is 0, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. **We call the number $n-1$ the degrees of freedom (df).**

Properties of the Student's-t Distribution

- The graph for the Student's-t distribution is similar to the Standard Normal curve.
- The mean for the Student's-t distribution is 0 and the distribution is symmetric about 0.
- The Student's-t distribution has more probability in its tails than the Standard Normal distribution because the spread of the t distribution is greater than the spread of the Standard Normal. So the graph of the Student's-t distribution will be thicker in the tails and shorter in the center than the graph of the Standard Normal distribution.
- The exact shape of the Student's-t distribution depends on the "degrees of freedom". As the degrees of freedom increases, the graph Student's-t distribution becomes more like the graph of the Standard Normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed but it is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's-t probabilities. The TI-83,83+,84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's-t distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's-t distribution.) When using t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's-t table (See the Table of Contents **15. Tables**) gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's-t probabilities.**

The notation for the Student's-t distribution is (using T as the random variable) is

- $T \sim t_{\text{df}}$ where $\text{df} = n - 1$.
- For example, if we have a sample of size n=20 items, then we calculate the degrees of freedom as df=n−1=20−1=19 and we write the distribution as $T \sim t_{19}$

If the population standard deviation is not known, the **error bound for a population mean** is:

- $\text{EBM} = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$
- use $\text{df} = n - 1$ degrees of freedom
- s = sample standard deviation

The format for the confidence interval is:

$(x - \text{EBM}, x + \text{EBM})$.

The TI-83, 83+ and 84 calculators have a function that calculates the confidence interval directly. To get to it,
Press `STAT`
Arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or just press `8`).

Example:
Exercise:

Problem:

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+ and 84+ calculators.
8.6 9.4 7.9 6.8 8.3 7.3 9.2 9.6 8.7 11.4 10.3 5.4 8.1 5.5 6.9

Solution:

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses the Ti-83+ and Ti-84 calculators (Solution B).

Solution A
To find the confidence interval, you need the sample mean, x, and the EBM.

$x = 8.2267 \qquad s = 1.6722 \qquad n = 15$

$$\text{df} = 15 - 1 = 14$$

$$\text{CL} = 0.95 \quad \text{so} \quad \alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 \qquad t_{\frac{\alpha}{2}} = t_{.025}$$

The area to the right of $t_{.025}$ is 0.025 and the area to the left of $t_{.025}$ is 1−0.025=0.975

$t_{\frac{\alpha}{2}} = t_{.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$\text{EBM} = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

$$\text{EBM} = 2.14 \cdot \left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$$x - \text{EBM} = 8.2267 - 0.9240 = 7.3$$

$$x + \text{EBM} = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is **(7.30, 9.15)**.

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B
Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to `Data` and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter .95
Arrow down to `Calculate` and press `ENTER`.
The 95% confidence interval is (7.3006, 9.1527)

Note: When calculating the error bound, a probability table for the Student's-t distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

**With contributions from Roberta Bloom

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $\text{CL} = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Degrees of Freedom (df)

The number of objects in a sample that are free to vary.

Error Bound for a Population Mean (EBM)

The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Normal Distribution

A continuous random variable (RV) with pdf $\mathrm{f(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student's-**t** Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Confidence Interval for a Population Proportion
Confidence Interval for a Population Proportion is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within 3 percentage points. Often, election polls are calculated with 95% confidence. So, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 : $(0.40 - 0.03, 0.40 + 0.03)$.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound,** and the **confidence level** for a proportion is similar to that for the population mean. The formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying **distribution is binomial**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n = the number of trials and p = the probability of a success. To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable $P\prime$ (read "P prime") is that proportion,

$$P\prime = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\widehat{P}$, read "P hat".)

When n is large and p is not close to 0 or 1, we can use the **normal distribution** to approximate the binomial.

$$X \sim N\left(n \cdot p, \sqrt{n \cdot p \cdot q}\right)$$

If we divide the random variable by n, the mean by n, and the standard deviation by n, we get a normal distribution of proportions with $P\prime$, called the estimated proportion, as the random variable. (Recall that a proportion = the number of successes divided by n.)

$$\frac{X}{n} = P\prime \sim N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right)$$

Using algebra to simplify : $\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$

$P\prime$ **follows a normal distribution for proportions**: $P\prime \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$

The confidence interval has the form $(p\prime - \text{EBP}, p\prime + \text{EBP})$.

$p\prime = \frac{x}{n}$

$p\prime$ = the **estimated proportion** of successes ($p\prime$ is a **point estimate** for p, the true proportion)

x = the **number** of successes.

n = the size of the sample

The error bound for a proportion is

$$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p\prime \cdot q\prime}{n}} \qquad \text{where } q\prime = 1 - p\prime$$

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{p \cdot q}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p\prime \cdot q\prime}{n}}$ as the standard deviation, instead of $\sqrt{\frac{p \cdot q}{n}}$

However, in the error bound formula, the standard deviation is $\sqrt{\frac{p\prime \cdot q\prime}{n}}$.

In the error bound formula, the **sample proportions** $p\prime$ **and** $q\prime$ **are estimates of the unknown population proportions** p **and** q. The estimated proportions $p\prime$ and $q\prime$ are used because p and q are not known. $p\prime$ and $q\prime$ are calculated from the data. $p\prime$ is the estimated proportion of successes. $q\prime$ is the estimated proportion of failures.

The confidence interval can only be used if the number of successes $np\prime$ and the number of failures $nq\prime$ are both larger than 5.

Note: For the normal distribution of proportions, the z-score formula is as follows.

If $P\prime \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ then the z-score formula is $z = \frac{p\prime - p}{\sqrt{\frac{p \cdot q}{n}}}$

Example:
Exercise:

Problem:

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adults residents of this city who have cell phones.

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution:

Let X = the number of people in the sample who have cell phones. X is binomial. X ~ $B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find $p\prime$, $q\prime$, and EBP.

$n = 500$ x = the number of successes $= 421$

$p\prime = \frac{x}{n} = \frac{421}{500} = 0.842$

$p\prime = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$q\prime = 1 - p\prime = 1 - 0.842 = 0.158$

Since $\text{CL} = 0.95$, then
$\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$.

Then $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$

Use the TI-83, 83+ or 84+ calculator command invNorm(0.975,0,1) to find $z_{.025}$. Remember that the area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.96 \cdot \sqrt{\frac{(0.842) \cdot (0.158)}{500}} = 0.032$$

$$p' - \text{EBP} = 0.842 - 0.032 = 0.81$$

$$p' + \text{EBP} = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - \text{EBP}, p' + \text{EBP}) = (0.810, 0.874)$.

Interpretation
We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level
95% of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution:

Using a function of the TI-83, 83+ or 84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x and enter 421.
Arrow down to n and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.81003, 0.87397).

Example:
Exercise:

Problem:

For a class project, a political science student at a large university wants to estimate the percent of students that are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students that are registered voters and interpret the confidence interval.

Solution:

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution A
$x = 300$ and $n = 500$.

$p' = \frac{x}{n} = \frac{300}{500} = 0.600$

$q' = 1 - p' = 1 - 0.600 = 0.400$

Since $\text{CL} = 0.90$, then
$\alpha = 1 - \text{CL} = 1 - 0.90 = 0.10 \qquad \frac{\alpha}{2} = 0.05$.

$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

Use the TI-83, 83+ or 84+ calculator command invNorm(0.95,0,1) to find $z_{.05}$. Remember that the area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p\prime \cdot q\prime}{n}} = 1.645 \cdot \sqrt{\frac{(0.60)\cdot(0.40)}{500}} = 0.036$$

$$p\prime - \text{EBP} = 0.60 - 0.036 = 0.564$$

$$p\prime + \text{EBP} = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p\prime - \text{EBP}, p\prime + \text{EBP}) = (0.564, 0.636)$.

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Using a function of the TI-83, 83+ or 84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x and enter 300.
Arrow down to n and enter 500.
Arrow down to `C-Level` and enter .90.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.564, 0.636).

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

- $\mathrm{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\text{p'q'}}{n}}$
- Solving for n gives you an equation for the sample size.
- $n = \frac{{z_{\frac{\alpha}{2}}}^{2} \cdot \text{p'q'}}{\mathrm{EBP}^{2}}$

Example:
Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ that use text messaging on their cell phone. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ that use text messaging on their cell phone.

Solution
From the problem, we know that **EBP=0.03** (3%=0.03) and

$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$ because the confidence level is 90%
However, in order to find n , we need to know the estimated (sample) proportion p'. Remember that q'=1-p'. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because p'q'=(.5)(.5)=.25 results in the largest possible product. (Try other products: (.6)(.4)=.24; (.3)(.7)=.21; (.2)(.8)=.16 and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.
$n = \frac{z^{2}\text{p'q'}}{\mathrm{EBP}^{2}}$ gives $n = \frac{1.645^{2}(.5)(.5)}{.03^{2}}$ =751.7
Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the

estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50+ that use text messaging on their cell phone.
**With contributions from Roberta Bloom.

Glossary

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: X~ $B(n,p)$. The mean is $\mu = \mathrm{np}$ and the standard deviation is $\sigma = \sqrt{\mathrm{npq}}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $\mathrm{CL} = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Proportion(EBP)

The margin of error. Depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

Normal Distribution

A continuous random variable (RV) with pdf $\mathrm{f}(\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: X ~ $N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Summary of Formulas

Formula General form of a confidence interval

$$(\text{lower value}, \text{upper value}) = (\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

FormulaTo find the error bound when you know the confidence interval

$\text{error bound} = \text{upper value} - \text{point estimate}$ OR $\text{error bound} = \frac{\text{upper value} - \text{lower value}}{2}$

FormulaSingle Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\overline{x} - \text{EBM}, \overline{x} + \text{EBM})$.

FormulaSingle Population Mean, Unknown Standard Deviation, Student's-t Distribution

Use the Student's-t Distribution with degrees of freedom $\text{df} = n - 1$. $\text{EBM} = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

FormulaSingle Population Proportion, Normal Distribution

Use the Normal Distribution for a single population proportion $p' = \frac{x}{n}$

$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}}$ $p' + q' = 1$

The confidence interval has the format $(p' - \text{EBP}, p' + \text{EBP})$.

FormulaPoint Estimates

$\overline{x}$ is a point estimate for μ

p' is a point estimate for ρ

s is a point estimate for σ

Practice 1: Confidence Intervals for Means, Known Population Standard Deviation

Student Learning Outcomes

- The student will calculate confidence intervals for means when the population standard deviation is known.

Given

The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. (http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm

Let $X =$ the age of a Winter Foothill College student

Calculating the Confidence Interval

Exercise:

Problem: $x =$

Solution:

30.4

Exercise:

Problem: n=

Solution:

25

Exercise:

Problem: 15=(insert symbol here)

Solution:

σ

Exercise:

Problem: Define the Random Variable, X, in words.

X =

Solution:

the mean age of 25 randomly selected Winter Foothill students

Exercise:

Problem: What is x estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Solution:

yes

Exercise:

Problem:

As a result of your answer to (4), state the exact distribution to use when calculating the Confidence Interval.

Solution:

Normal

Explaining the Confidence Interval

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students.

Exercise:

Problem: How much area is in both tails (combined)? $\alpha =$ ________

Solution:

0.05

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2}$ $=$ ________

Solution:

0.025

Exercise:

Problem: Identify the following specifications:

- **a** lower limit =
- **b** upper limit =
- **c** error bound =

Solution:

- **a**24.52
- **b**36.28
- **c**5.88

Exercise:

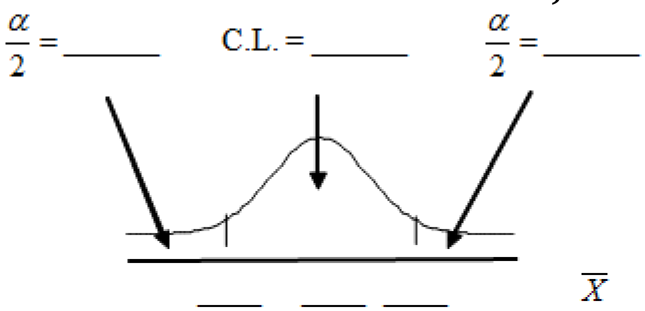
Problem: The 95% Confidence Interval is:__________________

Solution:

(24 52 36.28)

Exercise:

Problem:
Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval, and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:

Problem:

Using the same mean, standard deviation and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same mean, standard deviation and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Practice 2: Confidence Intervals for Means, Unknown Population Standard Deviation

Student Learning Outcomes

- The student will calculate confidence intervals for means when the population standard deviation is unknown.

Given

The following real data are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let $X =$ the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Calculating the Confidence Interval

Exercise:

Problem: Calculate the following:

- **a** $\overline{x} =$
- **b**$s_x =$
- **c**$n =$

Solution:

- **a**3.26
- **b**1.02
- **c**39

Exercise:

Problem:

Define the Random Variable, $\overline{X}$, in words. $\overline{X} =$ ______________________________

Solution:

the mean number of colors of 39 flags

Exercise:

Problem: What is $\overline{x}$ estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Solution:

No

Exercise:

Problem:

As a result of your answer to (4), state the exact distribution to use when calculating the Confidence Interval.

Solution:

t_{38}

Confidence Interval for the True Mean Number

Construct a 95% Confidence Interval for the true mean number of colors on national flags.

Exercise:

Problem: How much area is in both tails (combined)? $\alpha =$

Solution:

0.05

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2} =$

Solution:

0.025

Exercise:

Problem: Calculate the following:

- **a**lower limit =

- **b**upper limit =
- **c**error bound =

Solution:

- **a** 2.93
- **b**3.59
- **c**0.33

Exercise:

Problem: The 95% Confidence Interval is:

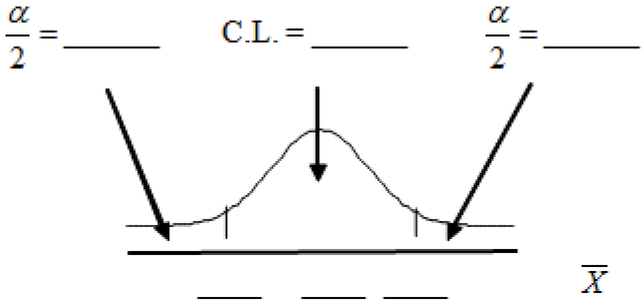
Solution:

2.93; 3.59

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:

Problem:

Using the same $\overline{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same $\overline{x}$, s_x, and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

Practice 3: Confidence Intervals for Proportions

Student Learning Outcomes

- The student will calculate confidence intervals for proportions.

Given

The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 - 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 - 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class is a random sample of the population.

Estimated Distribution

Exercise:

Problem: What is being counted?

Exercise:

Problem: In words, define the Random Variable X. $X =$

Solution:

The number of girls, age 8-12, in the beginning ice skating class

Exercise:

Problem: Calculate the following:

- **a** $x =$
- **b** $n =$
- **c** $p\prime=$

Solution:

- **a**64
- **b**80
- **c**0.8

Exercise:

Problem: State the estimated distribution of X. X ~

Solution:

$B(80, 0.80)$

Exercise:

Problem: Define a new Random Variable $P\prime$. What is $p\prime$ estimating?

Solution:

p

Exercise:

Problem: In words, define the Random Variable $P\prime$. $P\prime$=

Solution:

The proportion of girls, age 8-12, in the beginning ice skating class.

Exercise:

Problem: State the estimated distribution of $P\prime$. $P\prime$ ~

Explaining the Confidence Interval

Construct a 92% Confidence Interval for the true proportion of girls in the age 8 - 12 beginning ice-skating classes at the Ice Chalet.

Exercise:

Problem: How much area is in both tails (combined)? $\alpha =$

Solution:

1 - 0.92 = 0.08

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2} =$

Solution:

0.04

Exercise:

Problem: Calculate the following:

- **a**lower limit =
- **b**upper limit =
- **c**error bound =

Solution:

- **a**0.72
- **b**0.88
- **c**0.08

Exercise:

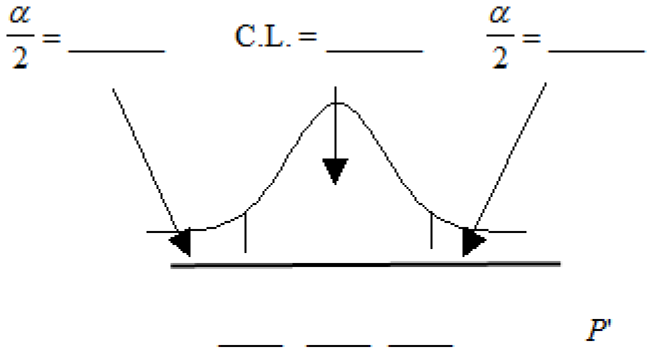
Problem: The 92% Confidence Interval is:

Solution:

(0.72; 0.88)

Exercise:

Problem:
Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval, and the sample proportion.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:

Problem:

Using the same $p\prime$ and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same $p\prime$ and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Exercise:

Problem:

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Homework

Note: If you are using a student's-t distribution for a homework problem below, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Exercise:

Problem:

Among various ethnic groups, the standard deviation of heights is known to be approximately 3 inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. 48 male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

- **a**
  - **i**$\overline{x}$ =________
  - **ii**σ = ________
  - **iii** s_x =________
  - **iv**n =________
  - **v**$n - 1$ =________
- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 95% confidence interval for the population mean height of male Swedes.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

- **e**What will happen to the level of confidence obtained if 1000 male Swedes are surveyed instead of 48? Why?

Solution:

- **a**
  - **i**71
  - **ii**3
  - **iii**2.8
  - **iv**48
  - **v**47
- **c**$N\left(71,\frac{3}{\sqrt{48}}\right)$
- **d**
  - **i**CI: (70.15,71.85)
  - **iii**EB = 0.85

Exercise:

Problem:

In six packages of “The Flintstones® Real Fruit Snacks” there were 5 Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

- **a**Define the Random Variables X and P’, in words.
- **b**Which distribution should you use for this problem? Explain your choice
- **c**Calculate p’.
- **d**Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
  - **i** State the confidence interval.
  - **ii**Sketch the graph.

    - **iii**Calculate the error bound.

- **e**Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

Exercise:

Problem:

A random survey of enrollment at 35 community colleges across the United States yielded the following figures (source: Microsoft Bookshelf): 6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622. Assume the underlying population is normal.

- **a**
    - **i**$\overline{x} =$
    - **ii** $s_x =$ ________
    - **iii**$n =$ ________
    - **iv**$n - 1 =$________
- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
    - **i**State the confidence interval.
    - **ii**Sketch the graph.
    - **iii**Calculate the error bound.
- **e**What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solution:

- **a**
  - **i**8629
  - **ii**6944
  - **iii**35
  - **iv**34
- **c** t_{34}
- **d**
  - **i**CI: (6244, 11,014)
  - **iii**EB = 2385
- **e**It will become smaller

Exercise:

Problem:

From a stack of IEEE Spectrum magazines, announcements for 84 upcoming engineering conferences were randomly picked. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

- **a**Define the Random Variables X and $\overline{X}$, in words.
- **b**Which distribution should you use for this problem? Explain your choice.
- **c**Construct a 95% confidence interval for the population mean length of engineering conferences.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

Exercise:

Problem:

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for service. The committee randomly surveyed 81 people. The sample mean was 8 hours with a sample standard deviation of 4 hours.

- **a**
  - **i**$\overline{x} =$________
  - **ii** $s_x =$ ________
  - **iii**$n =$________
  - **iv**$n - 1 =$________
- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 95% confidence interval for the population mean time wasted.
  - **a**State the confidence interval.
  - **b**Sketch the graph.
  - **c**Calculate the error bound.
- **e**Explain in a complete sentence what the confidence interval means.

Solution:

- **a**
  - **i**8
  - **ii**4
  - **iii**81
  - **iv**80

- **c** t_{80}
- **d**
  - **i**CI: (7.12, 8.88)
  - **iii**EB = 0.88

Exercise:

Problem:

Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

- **a**
  - **i**$\bar{x} =$ _________
  - **ii**$\sigma =$_________
  - **iii** $s_x =$_________
  - **iv**$n =$ _________
  - **v**$n - 1 =$_________
- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 90% confidence interval for the population mean time to complete the tax forms.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **e**If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

- **f**If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
- **g**Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within 1 hour. How would the number of people the firm surveys change? Why?

Exercise:

Problem:

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was 2 ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

- **a**
  - **i**$\overline{x} =$ _________
  - **ii**$\sigma =$ _________
  - **iii** $s_x =$_________
  - **iv**$n =$ _________
  - **v**$n - 1 =$ _________
- **b**Define the Random Variable X, in words.
- **c**Define the Random Variable $\overline{X}$, in words.
- **d**Which distribution should you use for this problem? Explain your choice.
- **e**Construct a 90% confidence interval for the population mean weight of the candies.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **f**Construct a 98% confidence interval for the population mean weight of the candies.

    - **i**State the confidence interval.
    - **ii**Sketch the graph.
    - **iii**Calculate the error bound.

- **g**In complete sentences, explain why the confidence interval in (f) is larger than the confidence interval in (e).
- **h**In complete sentences, give an interpretation of what the interval in (f) means.

Solution:

- **a**

    - **i**2
    - **ii**0.1
    - **iii** 0.12
    - **iv**16
    - **v**15

- **b**the weight of 1 small bag of candies
- **c**the mean weight of 16 small bags of candies
- **d**$N\left(2, \frac{0.1}{\sqrt{16}}\right)$
- **e**

    - **i** CI: (1.96, 2.04)
    - **iii** EB = 0.04

- **f**

    - **i** CI: (1.94, 2.06)
    - **iii** EB = 0.06

Exercise:

Problem:

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of 9 patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4 .

- **a**

  - **i**$\overline{x} =$ ________
  - **ii** $s_x =$ ________
  - **iii**$n =$ ________
  - **iv**$n - 1 =$ ________

- **b**Define the Random Variable X, in words.
- **c**Define the Random Variable $\overline{X}$, in words.
- **d**Which distribution should you use for this problem? Explain your choice.
- **e**Construct a 95% confidence interval for the population mean length of time.

  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

- **f**What does it mean to be “95% confident” in this problem?

Exercise:

Problem:

Suppose that 14 children were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of 6 months with a sample standard deviation of 3 months. Assume that the underlying population distribution is normal.

- **a**

- **i**$\overline{x}$ = __________
    - **ii** s_x = __________
    - **iii**n = __________
    - **iv**$n - 1$ = __________

- **b**Define the Random Variable X, in words.
- **c**Define the Random Variable $\overline{X}$, in words.
- **d**Which distribution should you use for this problem? Explain your choice.
- **e**Construct a 99% confidence interval for the population mean length of time using training wheels.
    - **i**State the confidence interval.
    - **ii**Sketch the graph.
    - **iii**Calculate the error bound.
- **f**Why would the error bound change if the confidence level was lowered to 90%?

Solution:

- **a**
    - **i**6
    - **ii**3
    - **iii**14
    - **iv**13
- **b**the time for a child to remove his training wheels
- **c**the mean time for 14 children to remove their training wheels.
- **d** t_{13}
- **e**
    - **i**CI: (3.58, 8.42)
    - **iii**EB = 2.42

Exercise:

Problem:

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

- **a**When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
- **b**If it was later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Exercise:

Problem:

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed to always buckle up. We are interested in the population proportion of drivers who claim to always buckle up.

- **a**
  - **i**$x =$ ________
  - **ii**$n =$ ________
  - **iii**$p' =$ ________
- **b**Define the Random Variables X and P', in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 95% confidence interval for the population proportion that claim to always buckle up.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

- **e**If this survey were done by telephone, list 3 difficulties the companies might have in obtaining random results.

Solution:

- **a**
  - **i**320
  - **ii** 400
  - **iii**0.80
- **c** $N\left(0.80, \sqrt{\frac{(0.80)(0.20)}{400}}\right)$
- **d**
  - **i** CI: (0.76, 0.84)
  - **iii** EB = 0.04

Exercise:

Problem:

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

- **a**
  - **i**$\overline{x} =$ _________
  - **ii** $s_x =$ _________
  - **iii**$n =$ _________
  - **iv**$n - 1 =$ _________
- **b**Define the Random Variables X and $\overline{X}$, in words.

- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii** Calculate the error bound.

Exercise:

Problem:

According to a recent survey of 1200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

- **a**Define the Random Variables X and P', in words.
- **b**Which distribution should you use for this problem? Explain your choice.
- **c**Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

Solution:

- **b** $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
- **c**
  - **i**CI: (0.59, 0.63)
  - **iii** EB = 0.02

Exercise:

Problem:

A survey of the mean amount of cents off that coupons give was done by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

- **a**
  - **i**$\overline{x} = \_\_\_\_\_\_\_\_\_$
  - **ii** $s_x = \_\_\_\_\_\_\_\_\_$
  - **iii**$n = \_\_\_\_\_\_\_\_\_$
  - **iv**$n - 1 = \_\_\_\_\_\_\_\_\_$
- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.
- **d**Construct a 95% confidence interval for the population mean worth of coupons.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **e**If many random samples were taken of size 14, what percent of the confident intervals constructed should contain the population mean worth of coupons? Explain why.

Exercise:

Problem:

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that their families would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

- **a**We are interested in finding the 95% confidence interval for the percent of all black families that would welcome a white person into their families. Define the Random Variables X and P', in words.
- **b**Which distribution should you use for this problem? Explain your choice.
- **c**Construct a 95% confidence interval
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.

Solution:

- **b** $N\left(0.86, \sqrt{\frac{(0.86)(0.14)}{323}}\right)$
- **c**
  - **i**CI: (0.823, 0.898)
  - **iii** EB = 0.038

Exercise:

Problem: Refer to the problem above.

- **a**Construct three 95% confidence intervals.

  - **i**Percent of all Asians that would welcome a white person into their families.
  - **ii**Percent of all Asians that would welcome a Latino into their families.
  - **iii**Percent of all Asians that would welcome a black person into their families.

- **b**Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
- **c**For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
- **d**For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Exercise:

Problem:

A camp director is interested in the mean number of letters each child sends during his/her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

- **a**

  - **i**$\overline{x} =$ _________
  - **ii**$\sigma =$ _________
  - **iii** $s_x =$ _________
  - **iv**$n =$ _________
  - **v**$n - 1 =$ _________

- **b**Define the Random Variables X and $\overline{X}$, in words.
- **c**Which distribution should you use for this problem? Explain your choice.

- **d**Construct a 90% confidence interval for the population mean number of letters campers send home.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **e**What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

Solution:

- **a**
  - **i** 7.9
  - **ii** 2.5
  - **iii** 2.8
  - **iv** 20
  - **v** 19
- **c** $N(7.9, \frac{2.5}{\sqrt{20}})$
- **d**
  - **i** CI: (6.98, 8.82)
  - **iii** EB: 0.92

Exercise:

Problem:

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight–year period.

- **a**Define the Random Variables X and P’, in words.

- **b**Which distribution should you use for this problem? Explain your choice.
- **c**Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **d** Explain what a “97% confidence interval” means for this study.

Exercise:

Problem:

In a recent sample of 84 used cars sales costs, the sample mean was $6425 with a standard deviation of $3156. Assume the underlying distribution is approximately normal.

- **a**Which distribution should you use for this problem? Explain your choice.
- **b**Define the Random Variable $\overline{X}$, in words.
- **c**Construct a 95% confidence interval for the population mean cost of a used car.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **d**Explain what a “95% confidence interval” means for this study.

Solution:

- **a** t_{83}
- **b**mean cost of 84 used cars
- **c**

- **i**CI: (5740.10, 7109.90)
- **iii** EB = 684.90

Exercise:

Problem:

A telephone poll of 1000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” 20% answered “crime”. We are interested in the population proportion of adult Americans who feel that crime is the main problem.

- **a**Define the Random Variables X and P', in words.
- **b**Which distribution should you use for this problem? Explain your choice.
- **c**Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **d**Suppose we want to lower the sampling error. What is one way to accomplish that?
- **e**The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3%. In 1-3 complete sentences, explain what the ± 3% represents.

Exercise:

Problem:

Refer to the above problem. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" 63% responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

1. Define the Random Variables X and P', in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population proportion of adult Americans worried a lot about the quality of education in our schools.

    - **i**State the confidence interval.
    - **ii**Sketch the graph.
    - **iii**Calculate the error bound.

4. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3%. In 1-3 complete sentences, explain what the ± 3% represents.

Solution:

- **b** $N\left(0.63, \sqrt{\frac{(0.63)(0.37)}{1000}}\right)$
- **c**

    - **i**CI: (0.60, 0.66)
    - **iii** EB = 0.03

Exercise:

Problem:

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

- **a**Calculate a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
  - **i**State the confidence interval.
  - **ii**Sketch the graph.
  - **iii**Calculate the error bound.
- **b**If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
- **c**Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
- **d**Calculate the mean.
- **e**Is the mean within the interval you calculated in part (a)? Did you expect it to be? Why or why not?

Exercise:

Problem:

A confidence interval for a proportion is given to be (– 0.22, 0.34). Why doesn’t the lower limit of the confidence interval make practical sense? How should it be changed? Why?

Try these multiple choice questions.

The next three problems refer to the following: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that “education and our schools” is one of the top issues facing

California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California. (Source: http://field.com/fieldpollonline/subscribers/)

Exercise:

Problem:A point estimate for the true population proportion is:

- **A**0.90
- **B**1.27
- **C**0.79
- **D**400

Solution:

C

Exercise:

Problem:A 90% confidence interval for the population proportion is:

- **A**(0.761, 0.820)
- **B**(0.125, 0.188)
- **C**(0.755, 0.826)
- **D**(0.130, 0.183)

Solution:

A

Exercise:

Problem: The error bound is approximately

- **A**1.581
- **B**0.791

- **C**0.059
- **D**0.030

Solution:

D

The next two problems refer to the following:

A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Exercise:

Problem:

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

- **A**(12.42, 14.18)
- **B**(12.32, 14.29)
- **C**(12.50, 14.10)
- **D**Impossible to determine

Solution:

B

Exercise:

Problem: What is the error bound?

- **A**0.87
- **B**1.98
- **C**0.99
- **D**1.74

Solution:

C

Exercise:

Problem:

What is meant by the term “90% confident” when constructing a confidence interval for a mean?

- **A**If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
- **B**If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
- **C**If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
- **D**If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Solution:

C

The next two problems refer to the following:

Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness and 338 did not.

Exercise:

Problem:

Find the Confidence Interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

- **A**(0.2975, 0.3796)
- **B**(0.6270, 6959)
- **C**(0.3041, 0.3730)
- **D**(0.6204, 0.7025)

Solution:

C

Exercise:

Problem:

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is:

- **A**0.6614
- **B**0.3386
- **C**173
- **D**338

Solution:

A

Review

The next three problems refer to the following situation: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to 3 pounds. Assume that the distribution for the weight loss is normal.

Exercise:

Problem:

To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

- **A** The number of people who lost weight on the special weight loss diet
- **B**The number of people who were on the diet
- **C**The mean amount of weight lost by 15 people on the special weight loss diet
- **D**The total amount of weight lost by 15 people on the special weight loss diet

Solution:

C

Exercise:

Problem: Find the probability asked for in the previous problem.

Solution:

0.9951

Exercise:

Problem:

Find the 90th percentile for the mean amount of weight lost by 15 people.

Solution:

12.99

The next three questions refer to the following situation: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

- So, if an accident occurs at 4 p.m., the amount of time, in hours, it took for the accident to occur is _______.
- $\mu =$ _______
- $\sigma^2 =$ _______

Exercise:

Problem:

What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

- **A**Cannot be determined from the information given
- **B** $\frac{1}{6}$
- **C** $\frac{1}{2}$
- **D** $\frac{1}{3}$

Solution:

C

Exercise:

Problem: The 20th percentile occurs after how many hours?

- **A** 0.20
- **B** 0.60
- **C** 0.50
- **D** 1

Solution:

B

Exercise:

Problem:

Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

- **A** $U(0,3)$
- **B** $\text{Exp}(\frac{1}{3})$
- **C** $N(60,5.477)$
- **D** $N(1.5,0.01875)$

Solution:

C

Exercise:

Problem:

Using the information in question #6, find the probability that the total time for all first accidents to occur is more than 43 hours.

Solution:

0.9990

The next two questions refer to the following situation: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from 1 to 15 days.

Exercise:

Problem:

How long must a parent expect to wait for his children to clean their rooms?

- **A** 8 days
- **B** 3 days
- **C** 14 days
- **D** 6 days

Solution:

A

Exercise:

Problem:

What is the probability that a parent will wait more than 6 days given that the parent has already waited more than 3 days?

- **A** 0.5174
- **B** 0.0174
- **C** 0.7500
- **D** 0.2143

Solution:

C

The next five problems refer to the following study: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

Exercise:

Problem:

Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

- **A** 80%
- **B** 20%
- **C** 30%
- **D** Cannot be determined

Solution:

A

Exercise:

Problem:

Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

- **A** 50%
- **B** 35%
- **C** 27.5%
- **D** 75%

Solution:

B

Exercise:

Problem:

Based upon the above information, are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

- **A** Yes
- **B** No
- **C** Cannot be determined

Solution:

B

Exercise:

Problem:

The interest rate charged on the financial aid is _______ data.

- **A** quantitative discrete
- **B** quantitative continuous
- **C**qualitative discrete
- **D** qualitative

Solution:

B

Exercise:

Problem:

What follows is information about the students who receive financial aid at the local community college.

- 1st quartile = $250

- 2nd quartile = $700
- 3rd quartile = $1200

(These amounts are for the school year.) If a sample of 200 students is taken, how many are expected to receive $250 or more?

- **A** 50
- **B** 250
- **C** 150
- **D** Cannot be determined

Solution:

- **C** 150

The next two problems refer to the following information: $P(A) = 0.2$, $P(B) = 0.3$, A and B are independent events.

Exercise:

Problem:$P(A \text{ AND } B) =$

- **A** 0.5
- **B** 0.6
- **C**0
- **D**0.06

Solution:

D

Exercise:

Problem:$P(A \text{ OR } B) =$

- **A** 0.56

- **B** 0.5
- **C**0.44
- **D**1

Solution:

C

Exercise:

Problem:

If H and D are mutually exclusive events, $P(H) = 0.25$, $P(D) = 0.15$, then $P(H|D)$

- **A** 1
- **B** 0
- **C** 0.40
- **D** 0.0375

Solution:

B

Introduction

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called **"hypothesis testing."** A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or fail to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note: To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See the Table of Contents topic "Solution Sheets".

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

Null and Alternate Hypotheses

The actual test begins by considering two hypotheses. They are called the **null hypothesis** and the **alternate hypothesis**. These hypotheses contain opposing viewpoints.

H_o: **The null hypothesis:** It is a statement about the population that will be assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt.

H_a: **The alternate hypothesis:** It is a claim about the population that is contradictory to H_o and what we conclude when we reject H_o.

Example:
H_o: No more than 30% of the registered voters in Santa Clara County voted in the primary election.
H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election.

Example:
We want to test whether the mean grade point average in American colleges is different from 2.0 (out of 4.0).
H_o: $\mu = 2.0$ H_a: $\mu \neq 2.0$

Example:
We want to test if college students take less than five years to graduate from college, on the average.
H_o: $\mu \geq 5$ H_a: $\mu < 5$

Example:

In an issue of **U. S. News and World Report**, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U. S. students take advanced placement exams and 4.4 % pass. Test if the percentage of U. S. students who take advanced placement exams is more than 6.6%.

H_o: p= 0.066 H_a: $p >$ 0.066

Since the null and alternate hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_o" if the sample information favors the alternate hypothesis or "do not reject H_o" or "fail to reject H_o" if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_o and H_a:

H_o	H_a
equal ($=$)	not equal ($\neq$) **or** greater than ($>$) **or** less than ($<$)
greater than or equal to ($\geq$)	less than ($<$)
less than or equal to ($\leq$)	more than ($>$)

Note:H_o always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use $=$ in the Null Hypothesis, even with $>$ or $<$ as the symbol in the Alternate Hypothesis. This practice is acceptable because we only make the decision to reject or not reject the Null Hypothesis.

Optional Collaborative Classroom Activity

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write a null and alternate hypotheses. Discuss your hypotheses with the rest of the class.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_o and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	**H_o IS ACTUALLY**	**...**
	True	False
Do not reject H_o	Correct Outcome	Type II error
Reject H_o	Type I Error	Correct Outcome

The four possible outcomes in the table are:

- The decision is to **not reject** H_o when, in fact, H_o **is true (correct decision).**
- The decision is to **reject** H_o when, in fact, H_o **is true** (incorrect decision known as a **Type I error**).
- The decision is to **not reject** H_o when, in fact, H_o **is false** (incorrect decision known as a **Type II error**).
- The decision is to **reject** H_o when, in fact, H_o **is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely 0.

The Power of the Test is $1 - \beta$. Ideally, we want a high power that is as close to 1 as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example:
Suppose the null hypothesis, H_o, is: Frank's rock climbing equipment is safe.
Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.
α **= probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β **= probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.
Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Example:
Suppose the null hypothesis, H_o, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.
Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.
α **= probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). β **= probability** that the

emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.
The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Glossary

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Type 2 Error

The decision is to not reject the Null hypothesis when, in fact, the Null hypothesis is false.

Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a **normal distribution** or a **student's-t distribution.** (Remember, use a student's-t distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.) In this chapter we perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$$X \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \qquad \text{or} \qquad t_{\text{df}}$$

The population parameter is μ. The estimated value (point estimate) for μ is x, the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

$$\text{P'} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is p. The estimated value (point estimate) for p is p'. p' $= \frac{x}{n}$ where x is the number of successes and n is the sample size.

Glossary

Normal Distribution

A continuous random variable (RV) with pdf $\text{f}(\text{x}) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student's-**t** Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Assumption

When you perform a **hypothesis test** **of a single population mean** μ using a **Student's-t distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{p \cdot q}{n}}$. Remember that $q = 1 - p$.

Glossary

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: X~ $B(n,p)$. The mean is $\mu = \mathrm{np}$ and the standard deviation is $\sigma = \sqrt{\mathrm{npq}}$. The probability of exactly x successes in n trials is $P\left(X = x\right) = \binom{n}{x}p^x q^{n-x}$.

Normal Distribution

A continuous random variable (RV) with pdf $\mathrm{f(x)} = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: X ~ $N(\mu,\sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student-**t** Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption** - it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Using the Sample to Test the Null Hypothesis

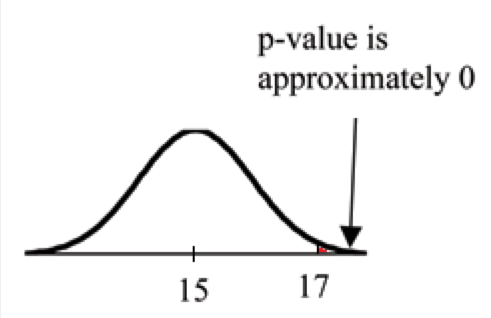
Use the sample data to calculate the actual probability of getting the test result, called the p-value. The p-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large p-value calculated from the data indicates that we should fail to reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example:
(to illustrate the p-value)
Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.
The null hypothesis could be H_o: $\mu \leq 15$ The alternate hypothesis is H_a: $\mu > 15$
The words **"is more than"** translates as a "$>$" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.
Since σ **is known** ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with mean μ= 15 and standard deviation $\frac{\sigma}{\sqrt{n}}$ $= \frac{0.5}{\sqrt{10}} = 0.16$.
Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the

sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.
The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means from Chapter 7.



p-value $= P(\overline{x} > 17)$ which is approximately 0.
A p-value of approximately 0 tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on the average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Glossary

Hypothesis
A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the **p-value** and a **preset or preconceived** α **(also called a "significance level")**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_o, do as follows:

- If $\alpha >$ p-value, reject H_o. The results of the sample data are significant. There is sufficient evidence to conclude that H_o is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If $\alpha \leq$ p-value, do not reject H_o. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.
- When you "do not reject H_o", it does not mean that you should believe that H_o is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Level of Significance of the Test

Probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Additional Information

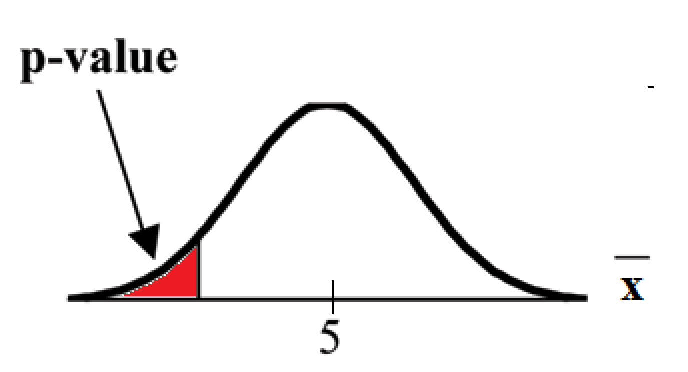
- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, the accepted standard is to use $\alpha = 0.05$.**
- When you calculate the **p-value** and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternate hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the p-value**: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value like 0.4, as opposed to a p-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in failing to reject the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left, right, and two-tailed test.

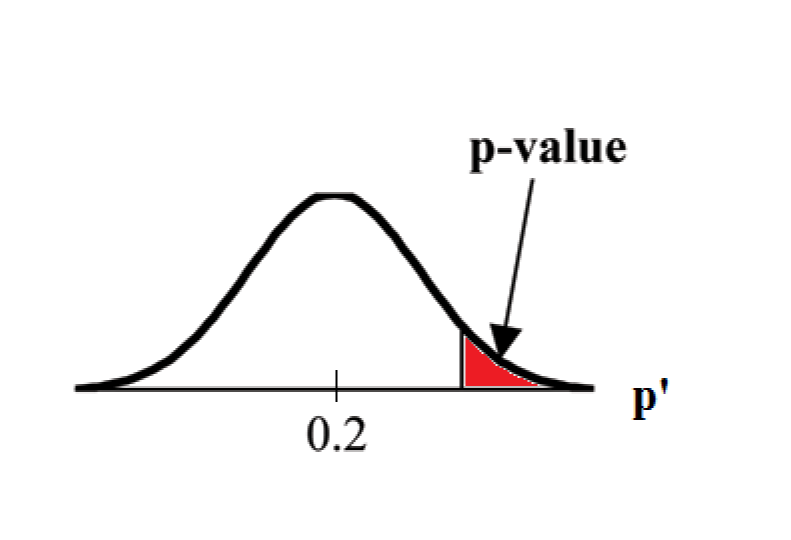
Example:

H_o: $\mu = 5$ H_a: $\mu < 5$

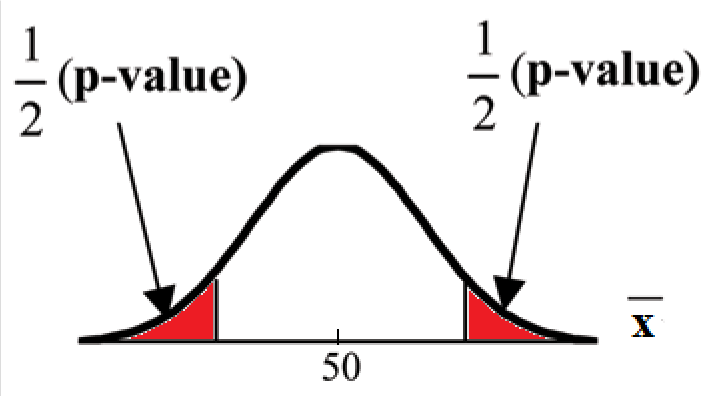
Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:



Example:
H_o: $p \leq 0.2$ H_a: $p > 0.2$
This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:



Example:
H_o: $\mu = 50$ H_a: $\mu \neq 50$
This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.



Glossary

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Summary of the Hypothesis Test

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_o and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the **p-value**. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_o), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

Glossary

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Examples

This module provides examples of Hypothesis Testing of a Single Mean and a Single Proportion as a part of the Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Example:

Exercise:

Problem:

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster by using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

H_o: $\mu = 16.43$ H_a: $\mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "$<$" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test: X is normal (population **standard deviation** is known: $\sigma = 0.8$)

$X \sim N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right)$ Therefore, $X \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

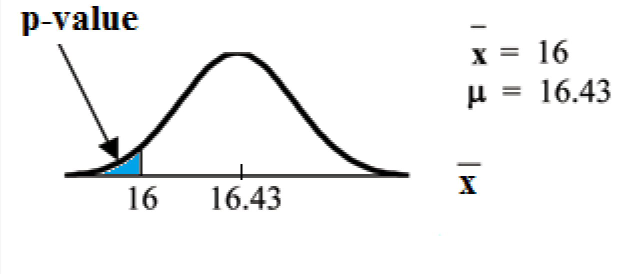
$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the p-value using the normal distribution for a mean:

$\text{p-value} = P\left(\bar{x} < 16\right) = 0.0187$ where the sample mean in the problem is given as 16.

$\text{p-value} = 0.0187$ (This is called the **actual level of significance**.) The p-value is the area to the left of the sample mean is given as 16.

Graph:



$\mu = 16.43$ comes from H_o. Our assumption is $\mu = 16.43$.

Interpretation of the p-value: If H_o is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the p-value:

$\alpha = 0.05$ $\text{p-value} = 0.0187$ $\alpha > \text{p-value}$

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ: (alternate hypothesis) and arrow over to $<\mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value ($p = 0.0187$) but it also calculates the test statistic (z-score) for the sample mean. $\mu < 16.43$ is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the p-value by doing a normal probability calculation using the Central Limit Theorem:

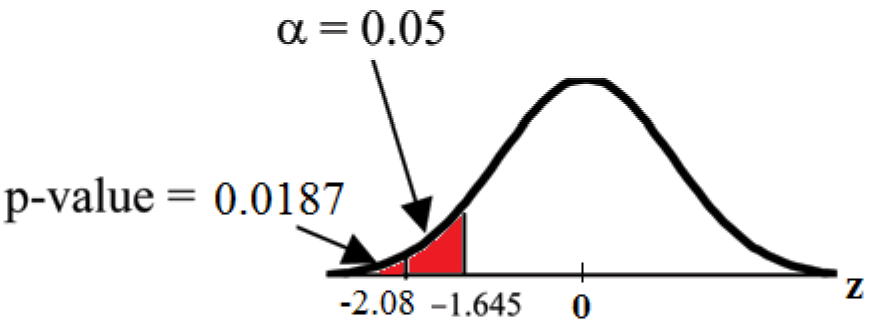
$P(x < 16\ ) =$ `2nd DISTR normcdf` $\left(-10\ \hat{}\ 99, 16, 16.43, 0.8/\sqrt{15}\right)$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Historical Note: The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is -2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{x-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$. For this problem, $x = 16$, $\mu_X = 16.43$ from the null hypothesis, $\sigma_X = 0.8$, and $n = 15$.) You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between -1.65 and -1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is -1.645. Since $-1.645 > -2.08$ (which demonstrates that $\alpha >$ p-value), reject H_o. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, and the test statistics and the critical value.



Example:
Exercise:

Problem:

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3) 215(3) 225(1) 241(2) 252(2) 265(2) 275(2) 313(2) 316(5) 338(2) 341(1) 345(2) 368(2) 385(1). (Source: data from Reuben Davis, Kraig Evans, and Scott Gunderson.)

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a **test of a single population mean**.

H_o: $\mu = 275$ H_a: $\mu > 275$ This is a right-tailed test.

Calculating the distribution needed:

Random variable: X = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

X ~ $N\left(275, \frac{55}{\sqrt{30}}\right)$

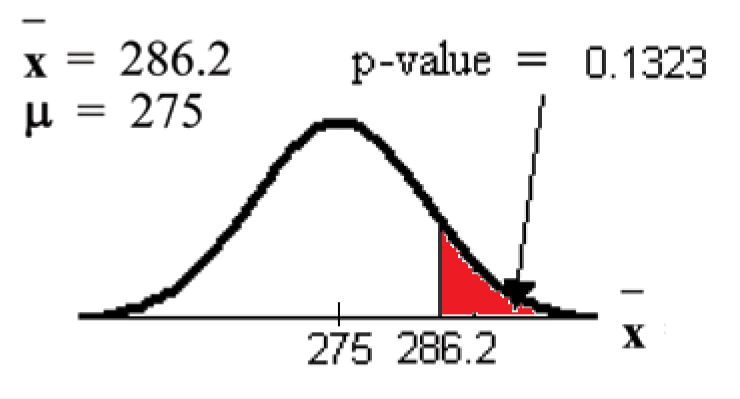
$x = 286.2$ pounds (from the data).

$\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see the calculator instructions below for using the data as input):

p-value $= P(\ x > 286.2) = 0.1323$.

Interpretation of the p-value: If H_o is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



Compare α and the p-value:

$\alpha = 0.025$ p-value $= 0.1323$

Make a decision: Since α<p-value, do not reject H_o.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`.

Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ : and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value ($p = 0.1331$, a little different from the above calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example:
Exercise:

Problem:

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are from a normal distribution.

Solution:

Set up the Hypothesis Test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

H_{o}: $\mu = 65$ $\qquad$ H_a: $\mu > 65$

Since the instructor thinks the average score is higher, use a "$>$". The "$>$" means the test is right-tailed.

Determine the distribution needed:

Random variable: X = average score on the first statistics test.

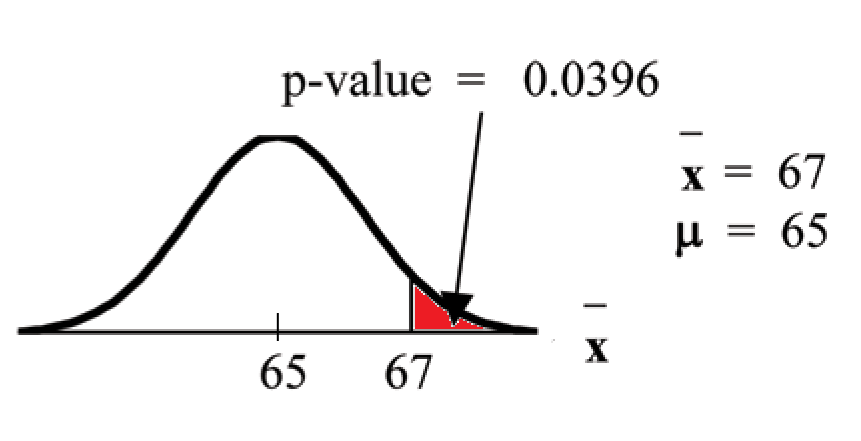
Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's-t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $\text{df} = 10 - 1 = 9$.

Calculate the p-value using the Student's-t distribution:

p-value $= P(\ x > 67\ ) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 67 or more.



Compare α and the p-value:

Since $\alpha = .05$ and p-value $= 0.0396$. Therefore, $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ : and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value ($p = 0.0396$) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example:
Exercise:

Problem:

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution:

Set up the Hypothesis Test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

H_o: $p = 0.50$ $\qquad$ H_a: $p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $P\prime$ = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for $P\prime$, the estimated proportion.

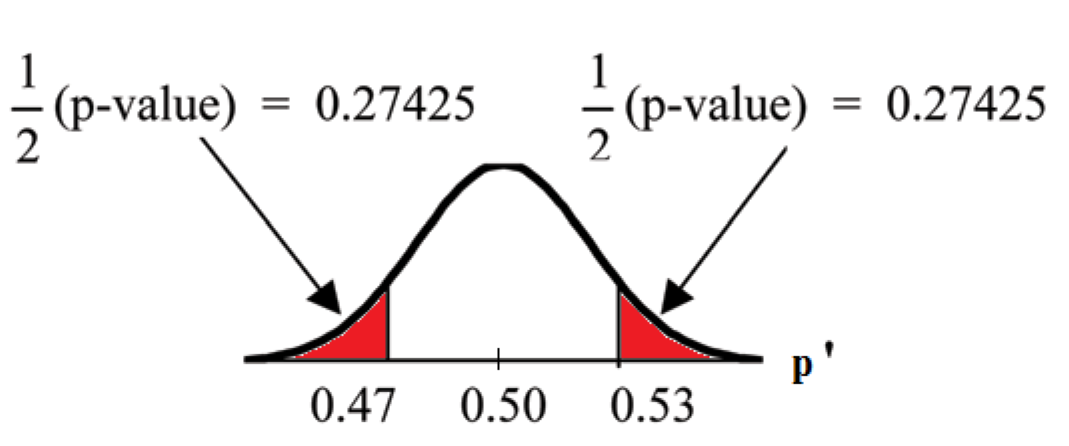
$P\prime \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ $\qquad$ Therefore, $P\prime \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$ where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$.

Calculate the p-value using the normal distribution for proportions:

p-value $= P($p'< 0.47 or p' > 0.53 $) = 0.5485$

where $x = 53$, $p\prime = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion $p\prime$ is 0.53 or more OR 0.47 or less (see the graph below).



$\mu = p = 0.50$ comes from H_o, the null hypothesis.

$p\prime = 0.53$. Since the curve is symmetrical and the test is two-tailed, the $p\prime$ for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value $= 0.5485$. Therefore, $\alpha <$ p-value.

Make a decision: Since $\alpha <$p-value, you cannot reject H_o.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides that are younger than their grooms is different from 50%.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press `STAT` and arrow over to `TESTS`. Press `5:1-PropZTest`. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press `ENTER`. Arrow down to

`Calculate` and press `ENTER`. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides that are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides that are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Example:
Exercise:

Problem:

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution:

Set up the Hypothesis Test:

H_o: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is $P\prime$ = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P\prime$ ~ N $\left(0.30, \sqrt{\frac{(0.30)\cdot(0.70)}{150}}\right)$

Exercise:

Problem:

The value that helps determine the p-value is $p\prime$. Calculate $p\prime$.

Solution:

$p\prime = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

p' = $\frac{43}{150}$

Exercise:

Problem: What is a **success** for this problem?

Solution:

A success is having three cell phones in a household.

Exercise:

Problem: What is the level of significance?

Solution:

The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Exercise:

Problem: Calculate the p-value.

Solution:

p-value = 0.7216

Exercise:

Problem:

Make a decision. ______________(Reject/Do not reject) H_0 because_____________.

Solution:

Assuming that α = 0.05, $\alpha <$ p-value. The Decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion $p\prime$ is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Note: Hypothesis testing problems consist of multiple steps. To help you do the problems, solution sheets are provided for your use. Look in the Table of Contents Appendix for the topic "Solution Sheets." If you like, use copies of the appropriate solution sheet for homework problems.

Example:

Exercise:

Problem:

My dog has so many fleas, They do not come off with ease. As for shampoo, I have tried many types Even one called Bubble Hype, Which only killed 25% of the fleas, Unfortunately I was not pleased. I've used all kinds of soap, Until I had give up hope Until one day I saw An ad that put me in awe. A shampoo used for dogs Called GOOD ENOUGH to Clean a Hog Guaranteed to kill more fleas. I gave Fido a bath And after doing the math His number of fleas Started dropping by 3's! Before his shampoo I counted 42. At the end of his bath, I redid the math And the new shampoo had killed 17 fleas. So now I was pleased. Now it is time for you to have some fun With the level of significance being .01, You must help me figure out Use the new shampoo or go without?

Solution:

Set up the Hypothesis Test:

$H_{\text{o}}: p = 0.25 \qquad H_{\text{a}}: p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable X or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

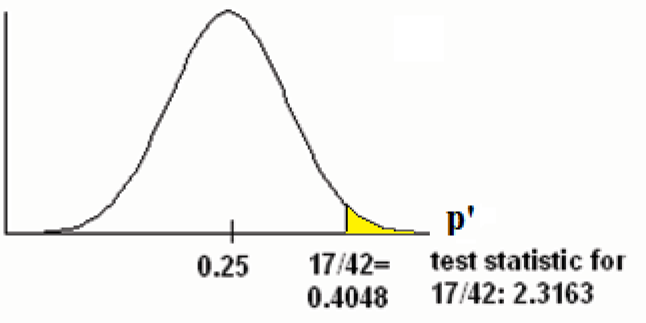
Calculate the p-value using the normal distribution for proportions:

p-value $=0.0103$

In 1 – 2 complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



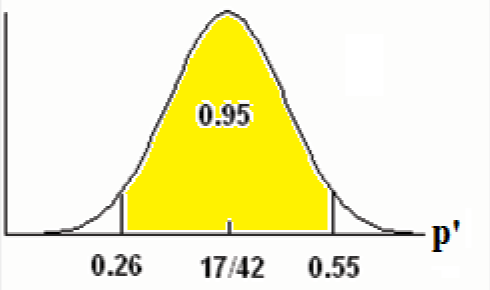
Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

alpha	decision	reason for decision
0.01	Do not reject H_o	α<p-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.



Confidence Interval: $(0.26, 0.55)$ We are 95% confident that the true population proportion p of fleas that are killed by the new shampoo is between 26% and 55%.

Note: This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\bar{X}\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(\mathrm{n\mu}, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Summary of Formulas

H_o and H_a are contradictory.

If H_o **has:**	equal	greater than or equal to	less than or equal to
then H_a **has:**	not equal **or** greater than **or** less than	less than	greater than

If α p-value, then do not reject H_o.

If α p-value, then reject H_o .

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

If there is no given preconceived α, then use α .

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's-t test**.
- Single population proportion: **Normal test**.

Practice 1: Single Mean, Known Population Standard Deviation
This module provides a practice of Hypothesis Testing of Single Mean and Single Proportion as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a hypothesis test of a single mean with known population standard deviation.

Given

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first–time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. The distribution of the population is normal.

Hypothesis Testing: Single Mean

Exercise:

Problem: Is this a test of means or proportions?

Solution:

Means

Exercise:

Problem: State the null and alternative hypotheses.

- **a**H_o:

- **b** H_a:

Solution:

- **a** H_o:$\mu = 2.5$ (or, H_o:$\mu \leq 2.5$)
- **b** H_a:$\mu > 2.5$

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

right-tailed

Exercise:

Problem: What symbol represents the Random Variable for this test?

Solution:

X

Exercise:

Problem: In words, define the Random Variable for this test.

Solution:

The mean time spent in jail for 26 first time convicted burglars

Exercise:

Problem:

Is the population standard deviation known and, if so, what is it?

Solution:

Yes, 1.5

Exercise:

Problem: Calculate the following:

- **a** $x =$
- **b** $\sigma =$
- **c** $s_x =$
- **d** $n =$

Solution:

- **a**3
- **b**1.5
- **c**1.8
- **d**26

Exercise:

Problem:

Since both σ and s_x are given, which should be used? In 1 -2 complete sentences, explain why.

Solution:

σ

Exercise:

Problem: State the distribution to use for the hypothesis test.

Solution:

$X \sim N(2.5, \frac{1.5}{\sqrt{26}})$

Exercise:

Problem:

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized mean and the sample mean x. Shade the area corresponding to the p-value.

Exercise:

Problem: Find the p-value.

Solution:

0.0446

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Solution:

- **a**Reject the null hypothesis

Discussion Questions

Exercise:

Problem:

Does it appear that the mean jail time spent for first time convicted burglars has increased? Why or why not?

Practice 2: Single Mean, Unknown Population Standard Deviation
This module provides a practice of Hypothesis Testing of Single Mean and Single Proportion as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a hypothesis test of a single mean with unknown population standard deviation.

Given

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

Hypothesis Testing: Single Mean

Exercise:

Problem: Is this a test of means or proportions?

Solution:

averages

Exercise:

Problem: State the null and alternative hypotheses.

- **a** H_o:
- **b** H_a:

Solution:

- **a** $H_o{:}\mu = 15$
- **b** $H_a{:}\mu \neq 15$

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

two-tailed

Exercise:

Problem: What symbol represents the Random Variable for this test?

Solution:

X

Exercise:

Problem: In words, define the Random Variable for this test.

Solution:

the mean time spent on death row for the 75 inmates

Exercise:

Problem:

Is the population standard deviation known and, if so, what is it?

Solution:

No

Exercise:

Problem: Calculate the following:

- **a** $x =$
- **b** $6.3 =$
- **c** $n =$

Solution:

- **a**17.4
- **b** s
- **c**75

Exercise:

Problem:

Which test should be used? In 1 -2 complete sentences, explain why.

Solution:

t $-$test

Exercise:

Problem: State the distribution to use for the hypothesis test.

Solution:

t_{74}

Exercise:

Problem:

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized mean and the sample mean, x. Shade the area corresponding to the p-value.

Exercise:

Problem: Find the p-value.

Solution:

0.0015

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Solution:

- **a**Reject the null hypothesis

Discussion Question

Does it appear that the mean time on death row could be 15 years? Why or why not?

Practice 3: Single Proportion
This module provides a practice of Hypothesis Testing of Single Mean and Single Proportion as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a hypothesis test of a single population proportion.

Given

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. (http://www.nimh.nih.gov/publicat/depression.cfm) Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

Hypothesis Testing: Single Proportion

Exercise:

Problem: Is this a test of means or proportions?

Solution:

Proportions

Exercise:

Problem: State the null and alternative hypotheses.

- **a**H_o:
- **b** H_a:

Solution:

- **a** H_o:$p = 0.095$
- **b** H_a:$p < 0.095$

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

left-tailed

Exercise:

Problem: What symbol represents the Random Variable for this test?

Solution:

P′

Exercise:

Problem: In words, define the Random Variable for this test.

Solution:

the proportion of people in that town surveyed suffering from depression or a depressive illness

Exercise:

Problem: Calculate the following:

- **a**$x =$
- **b** $n =$
- **c** p′=

Solution:

- **a**7
- **b**100
- **c**0.07

Exercise:

Problem:

Calculate $\sigma_{p'}$. Make sure to show how you set up the formula.

Solution:

0.0293

Exercise:

Problem: State the distribution to use for the hypothesis test.

Solution:

Normal

Exercise:

Problem:

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized mean and the sample proportion, p-hat. Shade the area corresponding to the p-value.

Exercise:

Problem: Find the p-value

Solution:

0.1969

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Solution:

- **a**Do not reject the null hypothesis

Discusion Question

Exercise:

Problem:

Does it appear that the proportion of people in that town with depression or a depressive illness is lower than general adult American population? Why or why not?

Homework

This module provides a homework of Hypothesis Testing of Single Mean and Single Proportion as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Exercise:

Problem:

Some of the statements below refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_o, and the alternative hypothesis, H_a, in terms of the appropriate parameter (μ or p).

- **a**The mean number of years Americans work before retiring is 34.
- **b**At most 60% of Americans vote in presidential elections.
- **c**The mean starting salary for San Jose State University graduates is at least $100,000 per year.
- **d**29% of high school seniors get drunk each month.
- **e**Fewer than 5% of adults ride the bus to work in Los Angeles.
- **f**The mean number of cars a person owns in her lifetime is not more than 10.
- **g**About half of Americans prefer to live away from cities, given the choice.
- **h**Europeans have a mean paid vacation each year of six weeks.
- **i**The chance of developing breast cancer is under 11% for women.
- **j**Private universities mean tuition cost is more than $20,000 per year.

Solution:

- **a** H_o:$\mu = 34$; H_a:$\mu \neq 34$
- **c** H_o:$\mu \geq 100{,}000$; H_a:$\mu < 100{,}000$
- **d** H_o:$p = 0.29$; H_a:$p \neq 0.29$
- **g** H_o:$p = 0.50$; H_a:$p \neq 0.50$
- **i** H_o:$p \geq 0.11$; H_a:$p < 0.11$

Exercise:

Problem:

For (a) - (j) above, state the Type I and Type II errors in complete sentences.

Solution:

- **a**Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We do not conclude that the mean is not 34 years, when it is not really 34 years.
- **c**Type I error: We conclude that the mean is less than \$100,000, when it really is at least \$100,000. Type II error: We do not conclude that the mean is less than \$100,000, when it is really less than \$100,000.
- **d**Type I error: We conclude that the proportion of h.s. seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We do not conclude that the proportion of h.s. seniors that get drunk each month is not 29%, when it is really not 29%.
- **i**Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We do not conclude that the proportion is less than 11%, when it really is less than 11%.

Exercise:

Problem: For (a) - (j) above, in complete sentences:

- **a**State a consequence of committing a Type I error.
- **b**State a consequence of committing a Type II error.

Note: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in 14. Appendix (online book version: the link is "Solution Sheets"; PDF book version: look under 14.5 Solution Sheets). Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note: If you are using a student's-t distribution for a homework problem below, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Exercise:

Problem:

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9800 miles. Do the data support the claim at the 5% level?

Exercise:

Problem:

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Solution:

- **e** $z = -2.71$
- **f**0.0034
- **h**Decision: Reject null; Conclusion: $\mu < 19$

- **i** $(17.449,18.757)$

Exercise:

Problem:

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

Exercise:

Problem:

An article in the **San Jose Mercury News** stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Solution:

- **e**3.5
- **f**0.0005
- **h**Decision: Reject null; Conclusion: $\mu > 4.5$
- **i** $(4.7553,5.4447)$

Exercise:

Problem:

The mean number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey 8 employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is about 10?

Exercise:

Problem:

In 1955, **Life Magazine** reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was 10. Does it appear that the mean work week has increased for women at the 5% level?

Solution:

- **e**2.7
- **f**0.0042
- **h**Decision: Reject Null
- **i** $(80.789,85.211)$

Exercise:

Problem:

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

Exercise:

Problem:

A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than 4. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Solution:

- **d** t_{11}
- **e**1.96
- **f**0.0380
- **h**Decision: Reject null when $a = 0.05$; do not reject null when $a = 0.01$
- **i** $(3.8865{,}5.9468)$

Exercise:

Problem:

Refer to the previous problem. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout’s mean I.Q. is **not** 4.

Exercise:

Problem:

According to an article in **Newsweek**, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Solution:

- **e**-1.64
- **f**0.1000
- **h**Decision: Do not reject null
- **i** $(0.3216, 0.4784)$

Exercise:

Problem:

A poll done for **Newsweek** found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only 2 had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the **Newsweek** poll? In complete sentences, also give three reasons why the two polls might give different results.

Exercise:

Problem:

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks 10 engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Solution:

- **d** t_9
- **e**-1.33
- **f**0.1086
- **h**Decision: Do not reject null
- **i** $(51.886,62.114)$

Exercise:

Problem:

Use the “Lap time” data for Lap 4 (see Table of Contents) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

Exercise:

Problem:

Use the “Initial Public Offering” data (see Table of Contents) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note: The following questions were written by past students. They are excellent problems!

Exercise:

Problem:18. "Asian Family Reunion" by Chau Nguyen

Every two years it comes around
We all get together from different towns.
In my honest opinion
It's not a typical family reunion
Not forty, or fifty, or sixty,
But how about seventy companions!
The kids would play, scream, and shout
One minute they're happy, another they'll pout.
The teenagers would look, stare, and compare
From how they look to what they wear.
The men would chat about their business
That they make more, but never less.
Money is always their subject
And there's always talk of more new projects.
The women get tired from all of the chats
They head to the kitchen to set out the mats.
Some would sit and some would stand
Eating and talking with plates in their hands.
Then come the games and the songs
And suddenly, everyone gets along!
With all that laughter, it's sad to say
That it always ends in the same old way.
They hug and kiss and say "good-bye"
And then they all begin to cry!
I say that 60 percent shed their tears
But my mom counted 35 people this year.
She said that boys and men will always have their pride,
So we won't ever see them cry.
I myself don't think she's correct,
So could you please try this problem to see if you object?

Exercise:

Problem:"The Problem with Angels" by Cyndy Dowling

```
Although this problem is wholly mine,
The catalyst came from the magazine, Time.
On the magazine cover I did find
The realm of angels tickling my mind.

Inside, 69% I found to be
In angels, Americans do believe.

Then, it was time to rise to the task,
Ninety-five high school and college students I
did ask.
Viewing all as one group,
Random sampling to get the scoop.

So, I asked each to be true,
"Do you believe in angels?"  Tell me, do!

Hypothesizing at the start,
Totally believing in my heart
That the proportion who said yes
Would be equal on this test.

Lo and behold, seventy-three did arrive,
Out of the sample of ninety-five.
Now your job has just begun,
Solve this problem and have some fun.
```

Solution:

- **e**1.65
- **f**0.0984
- **h**Decision: Do not reject null
- **i** $(0.6836, 0.8533)$

Exercise:

Problem:"Blowing Bubbles" by Sondra Prull

Studying stats just made me tense,
I had to find some sane defense.
Some light and lifting simple play
To float my math anxiety away.

Blowing bubbles lifts me high
Takes my troubles to the sky.
POIK! They're gone, with all my stress
Bubble therapy is the best.

The label said each time I blew
The average number of bubbles would be at least
22.
I blew and blew and this I found
From 64 blows, they all are round!

But the number of bubbles in 64 blows
Varied widely, this I know.
20 per blow became the mean
They deviated by 6, and not 16.

From counting bubbles, I sure did relax
But now I give to you your task.
Was 22 a reasonable guess?
Find the answer and pass this test!

Exercise:

Problem:21. "Dalmatian Darnation" by Kathy Sparling

A greedy dog breeder named Spreckles
Bred puppies with numerous freckles
The Dalmatians he sought

Possessed spot upon spot
The more spots, he thought, the more shekels.

His competitors did not agree
That freckles would increase the fee.
They said, “Spots are quite nice
But they don't affect price;
One should breed for improved pedigree.”

The breeders decided to prove
This strategy was a wrong move.
Breeding only for spots
Would wreak havoc, they thought.
His theory they want to disprove.

They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.

They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, “Well, shucks,
It's for one hundred one that I aim.”

Said an amateur statistician
Who wanted to help with this mission.
“Twenty-one for the sample
Standard deviation's ample:

They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle and dot
And tallied up every one.

```
Instead of one hundred one spots
They averaged ninety six dots
Can they muzzle Spreckles’
Obsession with freckles
Based on all the dog data they've got?
```

Solution:

- **e**-2.39
- **f**0.0093
- **h**Decision: Reject null
- **i** $(91.854, 100.15)$

Exercise:

Problem:

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:
Price per box of Mac and Cheese:

- 5 stores @ $2.02
- 15 stores @ $0.25

- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ 0.75.

I could see that the costs varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

Exercise:

Problem:

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark"
by Jacqueline Ghodsi
THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet’s tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter’s recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans’t thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn’d, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, “What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to’t I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether ‘tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

Solution:

- **e**-1.82
- **f**0.0345
- **h**Decision: Do not reject null
- **i** $(0.5331, 0.7685)$

Exercise:

Problem:"Untitled" by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

Exercise:

Problem: Japanese Girls’ Names

by Kumi Furuichi

It used to be very typical for Japanese girls’ names to end with “ko.” (The trend might have started around my grandmothers’ generation and its peak might have been around my mother’s generation.) “Ko” means “child” in Chinese character. Parents would name their daughters with “ko” attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko – a happy child, Yoshiko – a good child, Yasuko – a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with “ko.” More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother’s generation would have names with “ko” at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends’, ex-classmates’, co-workers, and acquaintances’ names that I could remember. Below are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi,

Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

Solution:

- **e** $z = -2.99$
- **f**0.0014
- **h**Decision: Reject null; Conclusion: $p < .70$
- **i** $(0.4529, 0.6582)$

Exercise:

Problem: Phillip’s Wish by Suzanne Osorio

```
My nephew likes to play
Chasing the girls makes his day.
He asked his mother
If it is okay
To get his ear pierced.
She said, “No way!”
To poke a hole through your ear,
Is not what I want for you, dear.
He argued his point quite well,
Says even my macho pal,  Mel,
Has gotten this done.
It’s all just for fun.
C’mon please, mom, please, what the hell.
Again Phillip complained to his mother,
Saying half his friends (including their
brothers)
```

Are piercing their ears
And they have no fears
He wants to be like the others.
She said, “I think it’s much less.
We must do a hypothesis test.
And if you are right,
I won’t put up a fight.
But, if not, then my case will rest.”
We proceeded to call fifty guys
To see whose prediction would fly.
Nineteen of the fifty
Said piercing was nifty
And earrings they’d occasionally buy.
Then there’s the other thirty-one,
Who said they’d never have this done.
So now this poem’s finished.
Will his hopes be diminished,
Or will my nephew have his fun?

Exercise:

Problem: The Craven by Mark Salangsang

Once upon a morning dreary
In stats class I was weak and weary.
Pondering over last night’s homework
Whose answers were now on the board
This I did and nothing more.

While I nodded nearly napping
Suddenly, there came a tapping.
As someone gently rapping,
Rapping my head as I snore.
Quoth the teacher, “Sleep no more.”

“In every class you fall asleep,”
The teacher said, his voice was deep.

“So a tally I’ve begun to keep
Of every class you nap and snore.
The percentage being forty-four.”

“My dear teacher I must confess,
While sleeping is what I do best.
The percentage, I think, must be less,
A percentage less than forty-four.”
This I said and nothing more.

“We’ll see,” he said and walked away,
And fifty classes from that day
He counted till the month of May
The classes in which I napped and snored.
The number he found was twenty-four.

At a significance level of 0.05,
Please tell me am I still alive?
Or did my grade just take a dive
Plunging down beneath the floor?
Upon thee I hereby implore.

Solution:

- **e**0.57
- **f**0.7156
- **h**Decision: Do not reject null
- **i** $(0.3415, 0.6185)$

Exercise:

Problem:

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%. (*Source: http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1*)

Exercise:

Problem:

68% of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC), CA, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents CA. NOTE: For more accurate results, use more CA community colleges and this past year's data. (Sources: **Growing by Degrees** by Allen and Seaman; Amit Schitai, Director of Instructional Technology and Distance Learning, LBCC).

Solution:

- **e**1.32
- **f**0.1873
- **h**Decision: Do not reject null
- **i** (0.65,0.90)

Exercise:

Problem:

According to an article in **Bloomberg Businessweek**, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year’s rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased. (*Source: http://www.businessweek.com/news/2011-09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html*)

Exercise:

Problem:

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test. (*Source: http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf*)

Solution:

- **e**9.98
- **f**0.0000
- **h**Decision: Reject null
- **i** (28.8,30.0)

Exercise:

Problem:

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test. (*Source: http://www.bls.gov/oes/current/oes291111.htm*)

Exercise:

Problem:

La Leche League International reports that the mean age of weaning a child from breastfeeding is age 4 to 5 worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was 9 months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than 4 years old. (*Source: http://www.lalecheleague.org/Law/BAFeb01.html*)

Solution:

- **e**-44.7
- **f**0.0000
- **h**Decision: Reject null
- **i** (0.60,0.90) - in years

Try these multiple choice questions.

Exercise:

Problem:

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

- **A**To conclude the drug is safe when in, fact, it is unsafe
- **B**To not conclude the drug is safe when, in fact, it is safe.
- **C**To conclude the drug is safe when, in fact, it is safe.
- **D**To not conclude the drug is unsafe when, in fact, it is unsafe

Solution:

B

The next two questions refer to the following information: Over the past few decades, public health officials have examined the link between weight concerns and teen girls smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three (63) said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?

Exercise:

Problem: The alternate hypothesis is

- **A** $p < 0.30$
- **B** $p \leq 0.30$
- **C** $p \geq 0.30$
- **D** $p > 0.30$

Solution:

D

Exercise:

Problem: After conducting the test, your decision and conclusion are

- **A**Reject H_o: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
- **B**Do not reject H_o: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
- **C**Do not reject H_o: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
- **D**Reject H_o: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

Solution:

C

The next three questions refer to the following information: A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of attended the midnight showing.

Exercise:

Problem: An appropriate alternative hypothesis is

- **A** $p = 0.20$
- **B** $p > 0.20$
- **C** $p < 0.20$
- **D** $p \leq 0.20$

Solution:

C

Exercise:

Problem: At a 1% level of significance, an appropriate conclusion is:

- **A**There is insufficient evidence to conclude that the percent of EVC students that attended the midnight showing of Harry Potter is less than 20%.
- **B**There is sufficient evidence to conclude that the percent of EVC students that attended the midnight showing of Harry Potter is more than 20%.
- **C**There is sufficient evidence to conclude that the percent of EVC students that attended the midnight showing of Harry Potter is less than 20%.

- **D**There is insufficient evidence to conclude that the percent of EVC students that attended the midnight showing of Harry Potter is at least 20%.

Solution:

A

Exercise:

Problem:

The Type I error is to conclude that the percent of EVC students who attended is

- **A**at least 20%, when in fact, it is less than 20%.
- **B**20%, when in fact, it is 20%.
- **C**less than 20%, when in fact, it is at least 20%.
- **D**less than 20%, when in fact, it is less than 20%.

Solution:

C

The next two questions refer to the following information:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than 7 hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than 7 hours of sleep per night, on average?

Exercise:

Problem: The distribution to be used for this test is X ~

- **A** $N(7.24,\frac{1.93}{\sqrt{22}})$
- **B** $N(7.24,1.93)$
- **C** t_{22}
- **D** t_{21}

Solution:

D

Exercise:

Problem:

The Type II error is to not reject that the mean number of hours of sleep LTCC students get per night is at least 7 when, in fact, the mean number of hours

- **A**is more than 7 hours.
- **B**is at most 7 hours.
- **C**is at least 7 hours.
- **D**is less than 7 hours.

Solution:

D

The next three questions refer to the following information: Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen (15) randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

Exercise:

Problem: The null and alternate hypotheses are:

- **A**H_o:$x = 4.5$, H_a:$x > 4.5$
- **B** H_o:$\mu \geq 4.5$ H_a:$\mu < 4.5$
- **C** H_o:$\mu = 4.75$ $H_{a:}\mu > 4.75$
- **D** H_o:$\mu = 4.5$ H_a:$\mu > 4.5$

Solution:

D

Exercise:

Problem:

At a significance level of $a = 0.05$, what is the correct conclusion?

- **A**There is enough evidence to conclude that the mean number of hours is more than 4.75
- **B**There is enough evidence to conclude that the mean number of hours is more than 4.5
- **C**There is not enough evidence to conclude that the mean number of hours is more than 4.5
- **D**There is not enough evidence to conclude that the mean number of hours is more than 4.75

Solution:

C

Exercise:

Problem: The Type I error is:

- **A**To conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher.
- **B**To conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same.

- **C**To conclude that the mean hours per week currently is 4.5, when in fact, it is higher.
- **D**To conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher.

Solution:

B

Review

This module provides an overview of Hypothesis Testing of Single Mean and Single Proportion as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Exercise:

Problem:

Rebecca and Matt are 14 year old twins. Matt’s height is 2 standard deviations below the mean for 14 year old boys’ height. Rebecca’s height is 0.10 standard deviations above the mean for 14 year old girls’ height. Interpret this.

- **A**Matt is 2.1 inches shorter than Rebecca
- **B**Rebecca is very tall compared to other 14 year old girls.
- **C**Rebecca is taller than Matt.
- **D**Matt is shorter than the average 14 year old boy.

Solution:

D

Exercise:

Problem:

Construct a histogram of the IPO data (see Table of Contents, 14. Appendix, Data Sets). Use 5 intervals.

Solution:

No solution provided. There are several ways in which the histogram could be constructed.

The next three exercises refer to the following information: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. X = the number of estimates.

x	**Rel. Freq.**	**Cumulative Rel. Freq.**
1	0.3	
2	0.2	
4	0.4	
5	0.1	

Complete the cumulative relative frequency column.

Exercise:

Problem:

Calculate the sample mean (a), the sample standard deviation (b) and the percent of the estimates that fall at or below 4 (c).

Solution:

- **a**2.8
- **b** 1.48
- **c**90%

Exercise:

Problem:

Calculate the median, M, the first quartile, Q1, the third quartile, Q3. Then construct a boxplot of the data.

Solution:

$M = 3$; $Q1 = 1$; $Q3 = 4$

Exercise:

Problem: The middle 50% of the data are between _____ and _____.

Solution:

1 and 4

The next three questions refer to the following table: Seventy 5th and 6th graders were asked their favorite dinner.

	Pizza	**Hamburgers**	**Spaghetti**	**Fried shrimp**
5th grader	15	6	9	0
6th grader	15	7	10	8

Exercise:

Problem:

Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

- **A** $\frac{32}{70}$
- **B** $\frac{8}{32}$
- **C** $\frac{8}{8}$
- **D** $\frac{8}{70}$

Solution:

D

Exercise:

Problem: Find the probability that a child does not prefer pizza.

- **A** $\frac{30}{70}$
- **B** $\frac{30}{40}$
- **C** $\frac{40}{70}$
- **D**1

Solution:

C

Exercise:

Problem:

Find the probability a child is in the 5th grade given that the child prefers spaghetti.

- **A** $\frac{9}{19}$
- **B** $\frac{9}{70}$
- **C** $\frac{9}{30}$
- **D** $\frac{19}{70}$

Solution:

A

Exercise:

Problem: A sample of convenience is a random sample.

- **A**true

- **B**false

Solution:

B

Exercise:

Problem: A statistic is a number that is a property of the population.

- **A**true
- **B**false

Solution:

B

Exercise:

Problem: You should always throw out any data that are outliers.

- **A**true
- **B**false

Solution:

B

Exercise:

Problem:

Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on the average. Of interest is the num.ber of pies she bakes each day

- **a**Define the Random Variable X.

- **b**State the distribution for X.
- **c**Find the probability that Lee bakes more than 25 pies in any given day.

Solution:

- **b** $P(20)$
- **c**0.1122

Exercise:

Problem:

Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, 5. Assume that the underlying distribution is normal. Calculate a 95% confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

Solution:

CI: $(5.52{,}8.48)$

Exercise:

Problem:

Given: uniform, exponential, normal distributions. Match each to a statement below.

- **a**mean = median ≠ mode
- **b**mean > median > mode
- **c**mean = median = mode

Solution:

- **a**uniform

- **b**exponential
- **c**normal

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Introduction

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported about various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

In the previous chapter, you learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** mean that the two samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

Note: This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and p-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using the TI-83+/TI-84 calculators, we do not need to separate two population means, independent groups, population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Becomes a test of one population mean.

Matched or Paired Samples
This module provides an overview of Hypothesis Testing: Matched or Paired Samples as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. The matched pairs have differences that either come from a population that is normal or the number of differences is sufficiently large so the distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student-t test for a single population mean with $n-1$ degrees of freedom where n is the number of differences.

Equation:

The test statistic (t-score) is:

$$t=\frac{x_d-\mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example:

Matched or paired samples

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution.

Subject:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Exercise:

Problem:

Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Solution:

Corresponding "before" and "after" values form matched pairs. (Calculate "sfter" - "before").

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data **for the test** are the differences: {0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6}

The sample mean and sample standard deviation of the differences are: $x_d = -3.13$ and $s_d = 2.91$ Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random Variable: X_d = the mean difference of the sensory measurements
Equation:

$$H_o : \mu_d \geq 0$$

There is no improvement. (μ_d is the population mean of the differences.)
Equation:

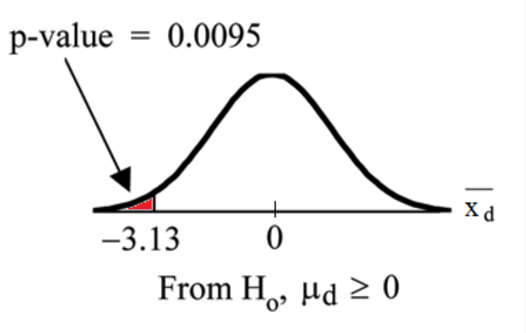
$$H_a : \mu_d < 0$$

There is improvement. The score should be lower after hypnotism so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a student-t with $\mathrm{df} = n - 1 = 8 - 1 = 7$. Use t_7**. (Notice that the test is for a single population mean.)**

Calculate the p-value using the Student-t distribution: $\text{p-value} = 0.0095$

Graph:



X_d is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$x_d = -3.13$

$s_d = 2.91$

Compare α and the p-value: $\alpha = 0.05$ and $\text{p-value} = 0.0095$. $\alpha > \text{p-value}$.

Make a decision: Since $\alpha > \text{p-value}$, reject H_o.

This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

Note: For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction and you will have the differences in the third list.

Note:TI-83+ and TI-84: Use your list of differences as the data. Press

```
STAT
```

and arrow over to

```
TESTS
```

. Press

```
2:T-Test
```

. Arrow over to

```
Data
```

and press

```
ENTER
```

. Arrow down and enter

```
0
```

for μ_0, the name of the list where you put the data, and

```
1
```

for Freq:. Arrow down to

```
μ
```

: and arrow over to

```
<
```

μ_0. Press

```
ENTER
```

. Arrow down to

```
Calculate
```

and press

```
ENTER
```

. The p-value is 0.0094 and the test statistic is -3.04. Do these instructions again except arrow to

```
Draw
```

(instead of

```
Calculate
```

). Press

```
ENTER
```

.

Example:
A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked 4 of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	**Player 1**	**Player 2**	**Player 3**	**Player 4**
Amount of weighted lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.
Exercise:

Problem:

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$x_d = 21.3$ $s_d = 46.7$

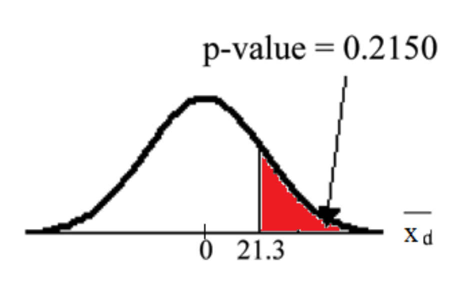
Using the difference data, this becomes a test of a single __________ (fill in the blank).

Define the random variable: $X_d =$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

$H_o : \mu_d \leq 0$ $H_a : \mu_d > 0$

Graph:



Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is to not reject the null hypothesis because $\alpha <$ p-value.

What is the conclusion?

Solution:

means; At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Example:

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The following data was collected.

Distance (in feet) using	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**	**Student 7**
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Exercise:

Problem:

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Note:use a t-test on the difference data. Assume the differences have a normal distribution. The random variable is the mean difference.

Note: The test statistic is 2.18 and the p-value is 0.0716.

What is your conclusion?

Solution:

H_0: μ_d equals 0; H_a: μ_d does not equal 0; Do not reject the null; At a 5% significance level, from the sample data, there is not sufficient evidence to conclude that the mean difference in distances between the children's dominant versus weaker hands is significant (there is not sufficient evidence to show that the children could push the shot-put further with their dominant hand). Alpha and the p-value are close so the test is not strong.

Comparing Two Independent Population Means with Unknown Population Standard Deviations

This module provides an overview of Comparing Two Independent Population Means with Unknown Population Standard Deviations as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

1. The two independent samples are simple random samples from two distinct populations.
2. Both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30. In that case, the populations need not be normally distributed.

Note: The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\overline{X_1}$ - $\overline{X_2}$, and divide by the standard error (shown below) in order to standardize the difference. The result is a t-score test statistic (shown below).

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\overline{X_1}$ - $\overline{X_2}$.

Equation:

The standard error is:

$$\sqrt{\frac{(S_1)^2}{n_1}+\frac{(S_2)^2}{n_2}}$$

The test statistic (t-score) is calculated as follows:

Equation:

t-score

$$\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{(S_1)^2}{n_1}+\frac{(S_2)^2}{n_2}}}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\overline{x_1}$ and $\overline{x_2}$ are the sample means. μ_1 and μ_2 are the population means.

The **degrees of freedom (df)** is a somewhat complicated calculation. However, a computer or calculator calculates it easily. The dfs are not always a whole number. The test statistic calculated above is approximated by the student's-t distribution with dfs as follows:

Equation:

Degrees of freedom

$$\text{df}=\frac{\left[\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}\right]^2}{\frac{1}{n_1-1}\cdot\left[\frac{(s_1)^2}{n_1}\right]^2+\frac{1}{n_2-1}\cdot\left[\frac{(s_2)^2}{n_2}\right]^2}$$

When both sample sizes n_1 and n_2 are five or larger, the student's-t approximation is very good. Notice that the sample variances ${s_1}^2$ and ${s_2}^2$ are not pooled. (If the question comes up, do not pool the variances.)

Note: It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example:
Independent groups
The average amount of time boys and girls ages 7 through 11 spend playing sports each day is believed to be the same. An experiment is done, data is collected, resulting in the table below. Both populations have a normal distribution.

	Sample Size	**Average Number of Hours Playing Sports Per Day**	**Sample Standard Deviation**
Girls	9	2 hours	$\sqrt{0.75}$
Boys	16	3.2 hours	1.00

Exercise:

Problem:

Is there a difference in the mean amount of time boys and girls ages 7 through 11 play sports each day? Test at the 5% level of significance.

Solution:

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{X_g} - \overline{X_b}$ = difference in the sample mean amount of time girls and boys play sports each day.

H_o: $\mu_g = \mu_b$ $\qquad\qquad$ $\mu_g - \mu_b = 0$

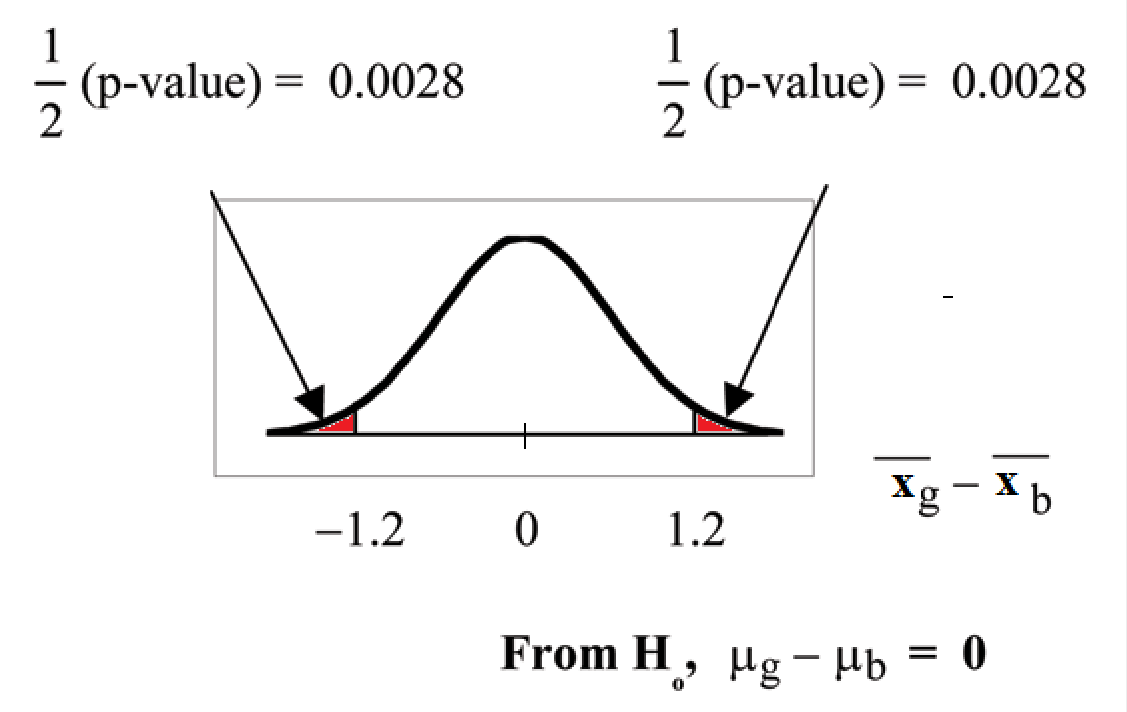
H_a: $\mu_g \neq \mu_b$ $\qquad\qquad$ $\mu_g - \mu_b \neq 0$

The words **"the same"** tell you H_o has an "=". Since there are no other words to indicate H_a, then assume **"is different."** This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the p-value using a student's-t distribution: p-value = 0.0054

Graph:



$s_g = \sqrt{0.75}$

$s_b = 1$

So, $\overline{x_g} - \overline{x_b} = 2 - 3.2 = -1.2$

Half the p-value is below -1.2 and half is above 1.2.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means you reject $\mu_g = \mu_b$. The means are different.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged 7 through 11 play sports per day is different (mean number of hours boys aged 7 through 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged 7 through 11 play sports per day is greater than the mean number of hours played by boys).

Note:TI-83+ and TI-84: Press

`STAT`

. Arrow over to

`TESTS`

and press

`4:2-SampTTest`

. Arrow over to Stats and press

`ENTER`

. Arrow down and enter

2

for the first sample mean,

$\sqrt{0.75}$

for Sx1,

9

for n1,

3.2

for the second sample mean,

1

for Sx2, and

16

for n2. Arrow down to μ1: and arrow to

does not equal

μ2. Press

ENTER

. Arrow down to Pooled: and

No

. Press

ENTER

. Arrow down to

`Calculate`

and press

`ENTER`

. The p-value is p = 0.0054, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Example:
A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples 9 graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from college A **has taken more math classes,** on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

Exercise:

Problem: Is this a test of two means or two proportions?

Solution:

two means

Exercise:

Problem:

Are the populations standard deviations known or unknown?

Solution:

unknown

Exercise:

Problem: Which distribution do you use to perform the test?

Solution:

student's-t

Exercise:

Problem: What is the random variable?

Solution:

$\overline{X_A} - \overline{X_B}$

Exercise:

Problem: What are the null and alternate hypothesis?

Solution:

- $H_o : \mu_A \leq \mu_B$
- $H_a : \mu_A > \mu_B$

Exercise:

Problem: Is this test right, left, or two tailed?

Solution:

right

Exercise:

Problem: What is the p-value?

Solution:

0.1928

Exercise:

Problem: Do you reject or not reject the null hypothesis?

Solution:

Do not reject.

Conclusion:
At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

Glossary

Degrees of Freedom (df)
The number of objects in a sample that are free to vary.

Standard Deviation
A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variable (Random Variable)

A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

Comparing Two Independent Population Proportions

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five and the number of failures is at least five for each of the samples.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions $\left(P'_A - P'_B\right)$ reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_o : p_A = p_B$. To conduct the test, we use a pooled proportion, p_c.

Equation:

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

Equation:

The distribution for the differences is:

$$P\prime_A - P\prime_B \text{~} N\left[0, \sqrt{p_c \cdot (1 - p_c) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

Equation:

The test statistic (z-score) is:

$$z = \frac{(p\prime_A - p\prime_B) - (p_A - p_B)}{\sqrt{p_c \cdot (1 - p_c) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Example:
Two population proportions
Two types of medication for hives are being tested to determine if there is a **difference in the proportions of adult patient reactions. Twenty** out of a random **sample of 200** adults given medication A still had hives 30 minutes after taking the medication. **Twelve** out of another **random sample of 200 adults** given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Determining the solution

This is a test of 2 population proportions.
Exercise:

Problem: How do you know?

Solution:

The problem asks for a difference in proportions.

Let A and B be the subscripts for medication A and medication B. Then p_A and p_B are the desired population proportions.

Random Variable:
$\text{P'}_A - \text{P'}_B =$ difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

$H_o: p_A = p_B \qquad\qquad p_A - p_B = 0$

$H_a: p_A \neq p_B \qquad\qquad p_A - p_B \neq 0$

The words **"is a difference"** tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 0.08 \quad 1 - p_c = 0.92$

Therefore,

$$\text{P'}_A - \text{P'}_B \text{~} N\left[0, \sqrt{(0.08) \cdot (0.92) \cdot \left(\frac{1}{200} + \frac{1}{200}\right)}\right]$$

$\text{P'}_A - \text{P'}_B$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution: p-value = 0.1404.

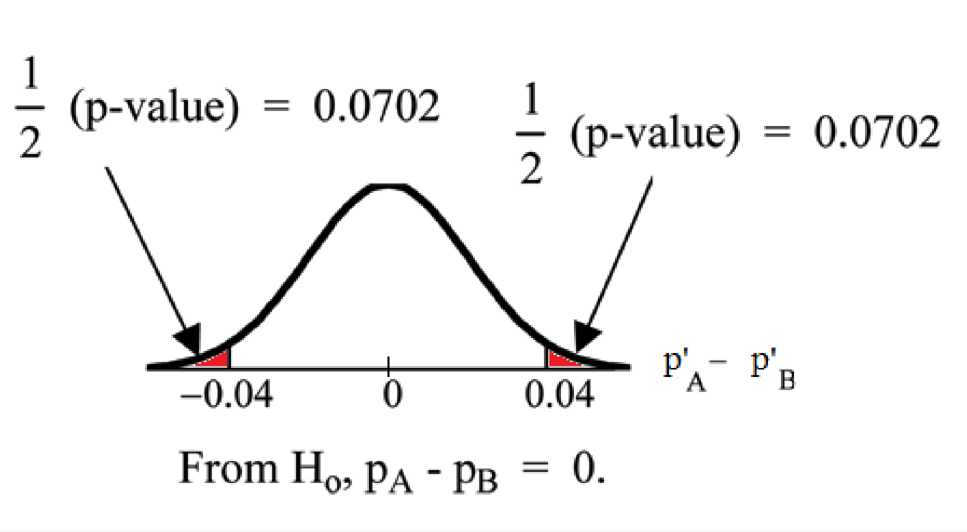
Estimated proportion for group A:

$$\text{p'}_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$$

Estimated proportion for group B:

$$\text{p'}_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$$

Graph:



$\text{P'}_A - \text{P'}_B = 0.1 - 0.06 = 0.04.$

Half the p-value is below -0.04 and half is above 0.04.

Compare α and the p-value: $\alpha = 0.01$ and the $\text{p-value} = 0.1404$. $\alpha < \text{p-value}$.

Make a decision: Since $\alpha < \text{p-value}$, do not reject H_o.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

Note:TI-83+ and TI-84: Press

`STAT`

. Arrow over to

`TESTS`

and press

`6:2-PropZTest`

. Arrow down and enter

`20`

for x1,

`200`

for n1,

`12`

for x2, and

`200`

for n2. Arrow down to

```
p1
```

: and arrow to

```
not equal p2
```

. Press

```
ENTER
```

. Arrow down to

```
Calculate
```

and press

```
ENTER
```

. The p-value is $p = 0.1404$ and the test statistic is 1.47. Do the procedure again but instead of

```
Calculate
```

do

```
Draw
```

.

Summary of Types of Hypothesis Tests

Two Population Means

- Populations are independent and population standard deviations are unknown.
- Populations are independent and population standard deviations are known (not likely).

Matched or Paired Samples

- Two samples are drawn from the same set of objects.
- Samples are dependent.

Two Population Proportions

- Populations are independent.

Homework

For questions [link] - [link], indicate which of the following choices best identifies the hypothesis test.

- **A**Independent group means, population standard deviations and/or variances known
- **B**Independent group means, population standard deviations and/or variances unknown
- **C**Matched or paired samples
- **D**Single mean
- **E**2 proportions
- **F**Single proportion

Exercise:

Problem:

A powder diet is tested on 49 people and a liquid diet is tested on 36 different people. The population standard deviations are 2 pounds and 3 pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

Solution:

A

Exercise:

Problem:

A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children that like the new chocolate bar is greater than the proportion of adults that like it.

Exercise:

Problem:

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 9 males and 16 females.

Solution:

B

Exercise:

Problem:

A football league reported that the mean number of touchdowns per game was 5. A study is done to determine if the mean number of touchdowns has decreased.

Exercise:

Problem:

A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. 100 students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively.

Solution:

A

Exercise:

Problem:

According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

Exercise:

Problem:

According to a recent study, U.S. companies have an mean maternity-leave of six weeks.

Solution:

D

Exercise:

Problem:

A recent drug survey showed an increase in use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.

Exercise:

Problem:

A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores.

Solution:

C

Exercise:

Problem:

University of Michigan researchers reported in the *Journal of the National Cancer Institute* that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.

Note: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in 14. Appendix (online book version: the link is "Solution Sheets"; PDF book version: look under 14.5 Solution Sheets). Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note: If you are using a student's-t distribution for a homework problem below, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Exercise:

Problem:

A powder diet is tested on 49 people and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had an mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had an mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Solution:

- **d** $t_{68.44}$
- **e**-1.04
- **f**0.1519
- **h**Decision: Do not reject null

Exercise:

Problem:

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of 3 English courses with a standard deviation of 0.8. The females took an average of 4 English courses with a standard deviation of 1.0. Are the means statistically the same?

Exercise:

Problem:

A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. 100 students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

Solution:

Standard Normal

- **e** $z = 2.14$
- **f**0.0163
- **h**Decision: Reject null when $\alpha = 0.05$; Do not reject null when $\alpha = 0.01$

Exercise:

Problem:

A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1200	1300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200

740	850

Exercise:

Problem:

A recent drug survey showed an increase in use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

Solution:

- **e**0.73
- **f**0.2326
- **h**Decision: Do not reject null

Exercise:

Problem:

A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5068 with a standard deviation of 4777. Of the 35 four-year colleges surveyed, the mean enrollment was 5466 with a standard deviation of 8191. (Source: *Microsoft Bookshelf*)

Exercise:

Problem:

A study was conducted by the U.S. Army to see if applying antiperspirant to soldiers' feet for a few days before a major hike would help cut down on the number of blisters soldiers had on their feet. In the experiment, for three nights before they went on a 13-mile hike, a group of 328 West Point cadets put an alcohol-based antiperspirant on their feet. A "control group" of 339 soldiers put on a similar, but inactive, preparation on their feet. On the day of the hike, the temperature reached 83° F. At the end of the hike, 21% of the soldiers who had used the antiperspirant and 48% of the control group had developed foot blisters. Conduct a hypothesis test to see if the proportion of soldiers using the antiperspirant was significantly lower than the control group. (Source: U.S. Army study reported in *Journal of the American Academy of Dermatologists*)

Solution:

- **e**-7.33
- **f**0
- **h**Decision: Reject null

Exercise:

Problem:

We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the white and the black races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4930. 580 were aged 15 to 24. The estimate for black females is 330. 40 were aged 15 to 24. We will let female suicide victims be our population. (Source*: the National Center for Health Statistics, U.S. Dept. of Health and Human Services*)

Exercise:

Problem:

At Rachel's 11th birthday party, 8 girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be 0. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Solution:

- **d** t_7
- **e**-1.51
- **f**0.1755
- **h**Decision: Do not reject null

Exercise:

Problem:

Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{l\text{arger}+\text{smaller dimension}}{l\text{arger dimension}}\right)$was the same in the Whitney Exhibit for works from 1900 – 1919 as for works from 1920 – 1942. 37 early works were sampled. They averaged 1.74 with a standard deviation of 0.11. 65 of the later works were sampled. They averaged 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation? (Source: *data from Whitney Exhibit on loan to San Jose Museum of Art*)

Exercise:

Problem:

One of the questions in a study of marital satisfaction of dual–career couples was to rate the statement, “I’m pleased with the way we divide the responsibilities for childcare.” The ratings went from 1 (strongly agree) to 5 (strongly disagree). Below are ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband’s versus the wife’s satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife’s score	2	2	3	3	4	2	1	1	2	4
Husband’s score	2	2	1	3	2	1	1	1	2	4

Solution:

- **d** t_9
- **e** $t = -1.86$
- **f**0.0479
- **h**Decision: Reject null, but run another test

Exercise:

Problem:

Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. Evaluate the data below. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Exercise:

Problem:

Mean entry level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. (Source: *http://* *www.graduatingengineer.com*). A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3450 and $4210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry level mechanical engineering salary is lower than the mean entry level electrical engineering salary.

Solution:

- **d** t_{108}

- **e** $t = -0.82$
- **f**0.2066
- **h**Decision: Do not reject null

Exercise:

Problem:

A recent year was randomly picked from 1985 to the present. In that year, there were 2051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different? (Source: *Chancellor's Office, California Community Colleges, November 1994*)

Exercise:

Problem:

Eight runners were convinced that the mean difference in their individual times for running one mile versus race walking one mile was at most 2 minutes. Below are their times. Do you agree that the mean difference is at most 2 minutes?

Running time (minutes)	Race walking time (minutes)
5.1	7.3
5.6	9.2
6.2	10.4
4.8	6.9
7.1	8.9
4.2	9.5

6.1	9.4
4.4	7.9

Solution:

- **d** t_7
- **e** $t = 2.9850$
- **f**0.0102
- **h**Decision: Reject null; There is sufficient evidence to conclude that the mean difference is more than 2 minutes.

Exercise:

Problem:

Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Exercise:

Problem:

While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. 8 of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Solution:

- **e**0.22
- **f**0.4133
- **h**Decision: Do not reject null

Exercise:

Problem:

We are interested in whether children’s educational computer software costs less, on average, than children’s entertainment software. 36 educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. 35 entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children’s educational software costs less, on average, than children’s entertainment software. (Source: *Educational Resources*, December catalog)

Exercise:

Problem:

Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

Solution:

- **e** $z = 2.50$
- **f**0.0063
- **h**Decision: Reject null

Exercise:

Problem:

A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1011, respectively. The population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

Exercise:

Problem:

Joan Nguyen recently claimed that the proportion of college–age males with at least one pierced ear is as high as the proportion of college–age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

Solution:

- **e**-4.82
- **f**0
- **h**Decision: Reject null

Exercise:

Problem:

Some manufacturers claim that non-hybrid sedan cars have a lower mean miles per gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of 7 mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of 4 mpg. Suppose that the population standard deviations are known to be 6 and 3, respectively. Conduct a hypothesis test to the manufacturers claim.

Questions [link] – [link] refer to the Terri Vogel’s data set (see Table of Contents).

Exercise:

Problem:

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

Solution:

- **d** $t_{20.32}$
- **e**-4.70
- **f**0.0001
- **h**Decision: Reject null

Exercise:

Problem: Repeat the test in [link], but use Lap 5 data this time.

Exercise:

Problem:

Repeat the test in [link], but this time combine the data from Laps 1 and 5.

Solution:

- **d** $t_{40.94}$
- **e**-5.08
- **f**0
- **h**Decision: Reject null

Exercise:

Problem:

In 2 – 3 complete sentences, explain in detail how you might use Terri Vogel’s data to answer the following question. “Does Terri Vogel drive faster in races than she does in practices?”

Exercise:

Problem:

Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

Solution:

- **e**-0.9223
- **f**0.1782
- **h**Decision: Do not reject null

Exercise:

Problem:"To Breakfast or Not to Breakfast?" by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, … birthdays. During this time, one looks back to see what he or she had achieved for the past year, and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, “This is why I always work more hours than you guys!”

And so, to prove his wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it is more than two hours. Using the data below, solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7

10	7
7	5
6	6
9	5

Try these multiple choice questions.

For questions [link] – [link], use the following information.

A new AIDS prevention drugs was tried on a group of 224 HIV positive patients. Forty-five (45) patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t= treated patient and ut= untreated patient.
Exercise:

Problem: The appropriate hypotheses are:

- **A** H_o:$p_t < p_{\text{ut}}$ and H_a:$p_t \geq p_{\text{ut}}$
- **B**H_o:$p_t \leq p_{\text{ut}}$ and H_a:$p_t > p_{\text{ut}}$
- **C** H_o:$p_t = p_{\text{ut}}$ and H_a:$p_t \neq p_{\text{ut}}$
- **D** H_o:$p_t = p_{\text{ut}}$ and H_a:$p_t < p_{\text{ut}}$

Solution:

D

Exercise:

Problem: If the p -value is 0.0062 what is the conclusion (use $\alpha = 0.05$)?

- **A**The method has no effect.

- **B**There is sufficient evidence to conclude that the method reduces the proportion of HIV positive patients that develop AIDS after four years.
- **C**There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients that develop AIDS after four years.
- **D**There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients that develop AIDS after four years.

Solution:

B

Exercise:

Problem:

Lesley E. Tan investigated the relationship between left-handedness and right-handedness and motor competence in preschool children. Random samples of 41 left-handers and 41 right-handers were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown below. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

- **A**Two independent means, normal distribution
- **B**Two independent means, student's-t distribution
- **C**Matched or paired samples, student's-t distribution
- **D**Two population proportions, normal distribution

Solution:

B

For questions [link] – [link], use the following information.

An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a “biofeedback exercise program.” Six (6) subjects were randomly selected and the blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated $(\text{after} - \text{before})$ producing the following results: $x_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training,

Exercise:

Problem: The distribution for the test is

- **A** t_5
- **B** t_6
- **C** $N(-10.2, 8.4)$
- **D** $N(-10.2, \frac{8.4}{\sqrt{6}})$

Solution:

A

Exercise:

Problem: If $\alpha = 0.05$, the p-value and the conclusion are

- **A**0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training
- **B**0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training
- **C**0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training
- **D**0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training

Solution:

C

For questions [link]– [link], use the following information.

The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to determine if the Western Reserve Division teams score, on average, fewer goals than the Eastern Reserve Division teams. Subscripts: **1** Western Reserve Division (**W**); **2** Eastern Reserve Division (**E**)
Exercise:

Problem: The **exact** distribution for the hypothesis test is:

- **A**The normal distribution.
- **B**The student's-t distribution.
- **C**The uniform distribution.
- **D**The exponential distribution.

Solution:

B

Exercise:

Problem: If the level of significance is 0.05, the conclusion is:

- **A**There is sufficient evidence to conclude that the **W** Division teams score, on average, fewer goals than the **E** teams.
- **B**There is insufficient evidence to conclude that the **W** Division teams score, on average, more goals than the **E** teams.
- **C**There is insufficient evidence to conclude that the **W** teams score, on average, fewer goals than the **E** teams score.
- **D**Unable to determine.

Solution:

C

Questions [link] – [link] refer to the following.

Neuroinvasive West Nile virus refers to a severe disease that affects a person’s nervous system . It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test. (Source: *http:// http://www.cdc.gov/ncidod/dvbid/westnile/index.htm*)

- “2011” subscript: 2011 group.
- “2010” subscript: 2010 group

Exercise:

Problem: This is:

- **A**a test of two proportions
- **B**a test of two independent means

- **C**a test of a single mean
- **D**a test of matched pairs.

Solution:

A

Exercise:

Problem: An appropriate null hypothesis is:

- **A** $p_{2011} \leq p_{2010}$
- **B** $p_{2011} \geq p_{2010}$
- **C** $\mu_{2011} \leq \mu_{2010}$
- **D** $p_{2011} > p_{2010}$

Solution:

A

Exercise:

Problem:

The p-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

- **A**There is sufficient evidence to conclude that the proportion of people in the United States in 2011 that got neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 that got neuroinvasive West Nile disease.
- **B**There is insufficient evidence to conclude that the proportion of people in the United States in 2011 that got neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 that got neuroinvasive West Nile disease.
- **C**There is insufficient evidence to conclude that the proportion of people in the United States in 2011 that got neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 that got neuroinvasive West Nile disease.
- **D**There is sufficient evidence to conclude that the proportion of people in the United States in 2011 that got neuroinvasive West Nile disease is

more than the proportion of people in the United States in 2010 that got neuroinvasive West Nile disease.

Solution:

D

Questions [link] and [link] refer to the following:

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-holes scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Exercise:

Problem: This is:

- **A**a test of two independent means
- **B**a test of two proportions
- **C**a test of a single proportion
- **D**a test of matched pairs.

Solution:

D

Exercise:

Problem: The correct decision is:

- **A**Reject H_o
- **B**Do not reject H_o

Solution:

B

Questions [link] and [link] refer to the following:

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students.

Exercise:

Problem: An appropriate alternate hypothesis for the hypothesis test is:

- **A** $\mu_{\text{day}} > \mu_{\text{night}}$
- **B** $\mu_{\text{day}} < \mu_{\text{night}}$
- **C** $\mu_{\text{day}} = \mu_{\text{night}}$
- **D** $\mu_{\text{day}} \neq \mu_{\text{night}}$

Solution:

D

Exercise:

Problem: A concluding statement is:

- **A**There is sufficient evidence to conclude that statistics night students mean on Exam 2 is better than the statistics day students mean on Exam

2.

- **B**There is insufficient evidence to conclude that the statistics day students mean on Exam 2 is better than the statistics night students mean on Exam 2.
- **C**There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
- **D**There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

Solution:

C

Review

The next three questions refer to the following information:

In a survey at Kirkwood Ski Resort the following information was recorded:

	0 – 10	11 - 20	21 - 40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Sport Participation by Age

Suppose that one person from of the above was randomly selected.

Exercise:

Problem:

Find the probability that the person was a skier or was age 11 – 20.

Solution:

$\frac{77}{100}$

Exercise:

Problem:

Find the probability that the person was a snowboarder given he/she was age 21 – 40.

Solution:

$\frac{12}{42}$

Exercise:

Problem: Explain which of the following are true and which are false.

- **a**Sport and Age are independent events.
- **b**Ski and age 11 – 20 are mutually exclusive events.
- **c** $P(\text{Ski and age } 21-40) < P(\text{Ski} \mid \text{age } 21-40)$
- **d** $P(\text{Snowboard or age } 0-10) < P(\text{Snowboard} \mid \text{age } 0-10)$

Solution:

- **a**False
- **b**False
- **c**True
- **d**False

Exercise:

Problem:

The average length of time a person with a broken leg wears a cast is approximately 6 weeks. The standard deviation is about 3 weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

Solution:

$N(180{,}16.43)$

Exercise:

Problem:

The distribution for X is Uniform. What can we say for certain about the distribution for X when $n = 1$?

- **A**The distribution for X is still Uniform with the same mean and standard dev. as the distribution for X.
- **B**The distribution for $X$is Normal with the different mean and a different standard deviation as the distribution for X.
- **C**The distribution for X is Normal with the same mean but a larger standard deviation than the distribution for X.
- **D**The distribution for X is Normal with the same mean but a smaller standard deviation than the distribution for X.

Solution:

A

Exercise:

Problem:

The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

- **A**The distribution for $\sum X$is still uniform with the same mean and standard deviation as the distribution for X.
- **B**The distribution for $\sum X$ is Normal with the same mean but a larger standard deviation as the distribution for X.
- **C**The distribution for $\sum X$ is Normal with a larger mean and a larger standard deviation than the distribution for X.
- **D**The distribution for $\sum X$ is Normal with the same mean but a smaller standard deviation than the distribution for X.

Solution:

C

The next three questions refer to the following information:

A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

Exercise:

Problem:

State the approximate distribution for X, the distribution for the average lengths of baby carrots in 16 five-pound bags. X~

Solution:

$N(2,\frac{.25}{\sqrt{16}})$

Exercise:

Problem:

Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

Exercise:

Problem: Find the probability that x is between 2 and 2.25 inches.

Solution:

0.5000

The next three questions refer to the following information:

At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of 5 minutes and a standard deviation of 2 minutes.

Exercise:

Problem: Find the 90th percentile of waiting time in minutes.

Solution:

7.6

Exercise:

Problem: Find the median waiting time for one student.

Solution:

5

Exercise:

Problem:

Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Solution:

0.9431

Introduction
This module provides an introduction to Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests.

Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to the above examples. This distribution is called the Chi-square distribution.

In this chapter, you will learn the three major applications of the Chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as with the lottery example
- The test of independence, which determines if events are independent, such as with the movie example

- The test of a single variance, which tests variability, such as with the coffee example

Note: Though the Chi-square calculations depend on calculators or computers for most of the calculations, there is a table available (see the Table of Contents **15. Tables**). TI-83+ and TI-84 calculator instructions are included in the text.

Optional Collaborative Classroom Activity

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, etc.). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

Notation

This module provides an overview of Chi-Square Distribution Notation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The notation for the chi-square distribution is:

χ ~ χ

where degrees of freedom depend on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use n . The degrees of freedom for the three major uses are each calculated differently.)

For the χ distribution, the population mean is μ and the population standard deviation is σ .

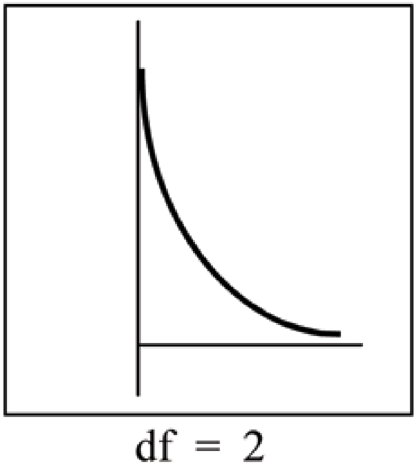
The random variable is shown as χ but may be any upper case letter.

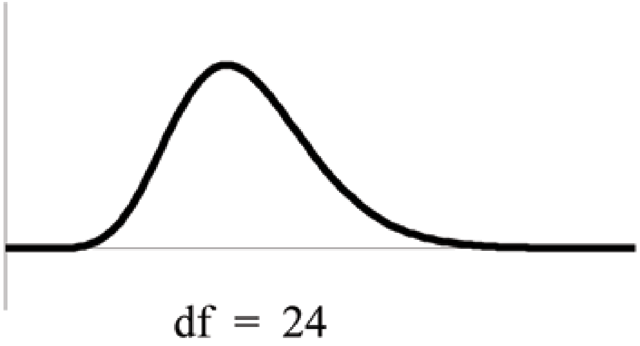
The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

χ Z Z Z_k

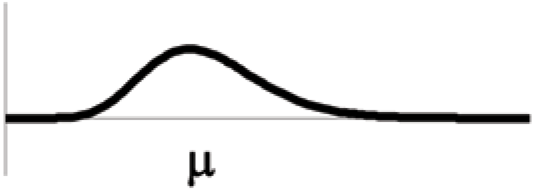
Facts About the Chi-Square Distribution

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.





3. The test statistic for any test is always greater than or equal to zero.
4. When $\text{df} > 90$, the chi-square curve approximates the normal. For X ~ χ^2_{1000} the mean, $\mu = \text{df} = 1000$ and the standard deviation, $\sigma = \sqrt{2 \cdot 1000} = 44.7$. Therefore, X ~ $N(1000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.



In the next sections, you will learn about four different applications of the Chi-Square Distribution. These hypothesis tests are almost always right-tailed tests. In order to understand why the tests are mostly right-tailed, you will need to look carefully at the actual definition of the test statistic. Think about the following while you study the next four sections. If the expected and observed values are "far" apart, then the test statistic will be "large" and we will reject in the right tail. The only way to obtain a test statistic very close to zero, would be if the observed and expected values are very, very close to each other. A left-tailed test could be used to determine if the fit were "too good." A "too good" fit might occur if data had been manipulated or invented. Think about the implications of right-tailed versus left-tailed hypothesis tests as you learn the applications of the Chi-Square Distribution.

Goodness-of-Fit Test

This module describes how the chi-square distribution is used to conduct goodness-of-fit test.

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:
Equation:

$$\sum_k \frac{(O-E)^2}{E}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The degrees of freedom are $\text{df} = (\text{number of categories - 1})$.

The goodness-of-fit test is almost always right tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note: The expected value for each cell needs to be at least 5 in order to use this test.

Example:
Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism follows faculty perception. The

faculty expected that a group of 100 students would miss class according to the following chart.

Number absences per term	**Expected number of students**
0 - 2	50
3 - 5	30
6 - 8	12
9 - 11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number **(observed)** of absences in a course. The next chart displays the result of that survey.

Number absences per term	**Actual number of students**
0 - 2	35
3 - 5	40
6 - 8	20
9 - 11	1
12+	4

Determine the null and alternate hypotheses needed to conduct a goodness-of-fit test.
H_o**:** Student absenteeism **fits** faculty perception.

The alternate hypothesis is the opposite of the null hypothesis.

H_a**:** Student absenteeism **does not fit** faculty perception.

Exercise:

Problem:

Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution:

No. Notice that the expected number of absences for the "12+" entry is less than 5 (it is 2). Combine that group with the "9 - 11" group to create new tables where the number of students for each entry are at least 5. The new tables are below.

Number absences per term	**Expected number of students**
0 - 2	50
3 - 5	30
6 - 8	12
9+	8

Number absences per term	**Actual number of students**
0 - 2	35
3 - 5	40
6 - 8	20
9+	5

Exercise:

Problem: What are the degrees of freedom (df)?

Solution:

There are 4 "cells" or categories in each of the new tables.

df = number of cells - 1 = 4 - 1 = 3

Example:
Employers particularly want to know which days of the week employees are absent in a five day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week did they have the highest number of employee absences. The results were distributed as follows:

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Number of Absences	15	12	9	9	15

Day of the Week Employees were most Absent

Exercise:

Problem:

For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five day work week? Test at a 5% significance level.

Solution:

The null and alternate hypotheses are:

- H_o: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- $(O - E)$
- $(O - E)^2$
- $\frac{(O-E)^2}{E}$

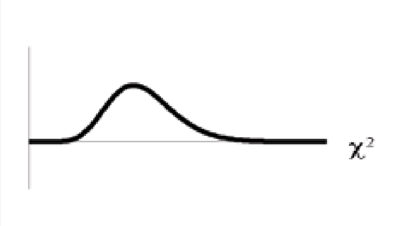
The last column ($\frac{(O-E)^2}{E}$) should have 0.75, 0, 0.75, 0.75, 0.75.
Now add (sum) the last column. Verify that the sum is 3. This is the χ^2 test statistic.

To find the p-value, calculate $P\left(\chi^2 > 3\right)$. This test is right-tailed.
(Use a computer or calculator to find the p-value. You should get $\text{p-value} = 0.5578$.)

The dfs are the number of cells $- 1 =$ 5 - 1 $= 4$.

TI-83+ and TI-84: Press `2nd DISTR`. Arrow down to χ^2`cdf`. Press `ENTER`. Enter `(3,10^99,4)`. Rounded to 4 decimal places, you should see 0.5578 which is the p-value.

Next, complete a graph like the one below with the proper labeling and shading. (You should shade the right tail.)



The decision is to not reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

Note:TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example (Example 11-3) has the calculator instructions. The newer TI-84 calculators have in

```
STAT TESTS
```

the test

```
Chi2 GOF
```

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

```
STAT
```

```
TESTS
```

and

```
Chi2 GOF
```

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

```
calculate
```

or

```
draw
```

. Make sure you clear any lists before you start. See below.

Note:To Clear Lists in the calculators: Go into

```
STAT EDIT
```

and arrow up to the list name area of the particular list. Press

`CLEAR`

and then arrow down. The list will be cleared. Or, you can press

`STAT`

and press 4 (for

`ClrList`

). Enter the list name and press

`ENTER`

.

Example:
One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Number of Televisions	**Percent**
0	10
1	16
2	55
3	11
over 3	8

The table contains expected (E) percents.
A random sample of 600 families in the far western United States resulted in the following data:

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
over 3	15
	Total = 600

The table contains observed (O) frequency values.

Exercise:

Problem:

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution:

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are:

Number of Televisions	**Percent**	**Expected Frequency**
0	10	$(0.10) \cdot (600) = 60$
1	16	$(0.16) \cdot (600) = 96$
2	55	$(0.55) \cdot (600) = 330$
3	11	$(0.11) \cdot (600) = 66$
over 3	8	$(0.08) \cdot (600) = 48$

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter .10*600.

H_o: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

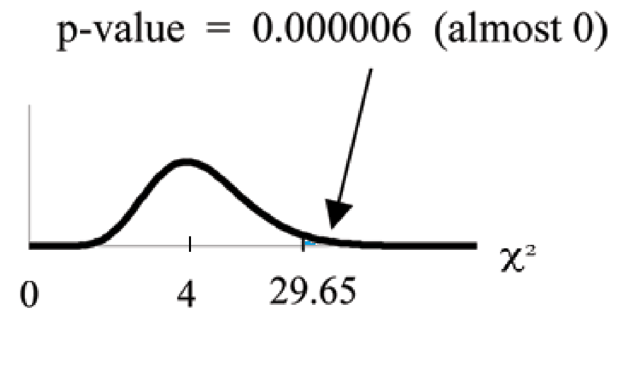
H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $\text{df} = (\text{the number of cells}) - 1 = 5 - 1 = 4$.

Note:$\text{df} \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:



Probability statement: p-value $= P(\chi^2 > 29.65) = 0.000006$.

Compare α and the p-value:

- $\alpha = 0.01$
- p-value $= 0.000006$

So, $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Note:TI-83+ and some TI-84 calculators: Press

`STAT`

and

`ENTER`

. Make sure to clear lists

`L1`

,

L2

, and

L3

if they have data in them (see the note at the end of Example 11-2). Into

L1

, put the observed frequencies

66

,

119

,

349

,

60

,

15

. Into

L2

, put the expected frequencies

.10*600, .16*600

,

.55*600

,

.11*600

,

`.08*600`

. Arrow over to list

`L3`

and up to the name area

`"L3"`

. Enter

`(L1-L2)^2/L2`

and

`ENTER`

. Press

`2nd QUIT`

. Press

`2nd LIST`

and arrow over to

`MATH`

. Press

`5`

. You should see

`"sum" (Enter L3)`

. Rounded to 2 decimal places, you should see

`29.65`

. Press

`2nd DISTR`

. Press

`7`

or Arrow down to

`7:χ2cdf`

and press

`ENTER`

. Enter

`(29.65,1E99,4)`

. Rounded to 4 places, you should see

`5.77E-6 = .000006`

(rounded to 6 decimal places) which is the p-value.

The newer TI-84 calculators have in

`STAT TESTS`

the test

`Chi2 GOF`

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

`STAT`

`TESTS`

and

`Chi2 GOF`

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

`calculate`

or

`draw`

. Make sure you clear any lists before you start.

Example:
Exercise:

Problem:

Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution:

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the value 0, 1, 2. (There are 0, 1, or 2 heads in the flip of 2 coins.) Therefore, the **number of cells is 3**. Since X = the number of heads, the observed frequencies are 20 (for 2 heads), 57 (for 1 head), and 23 (for 0 heads or both tails). The expected frequencies are 25 (for 2 heads), 50 (for 1 head), and 25 (for 0 heads or both tails). This test is right-tailed.

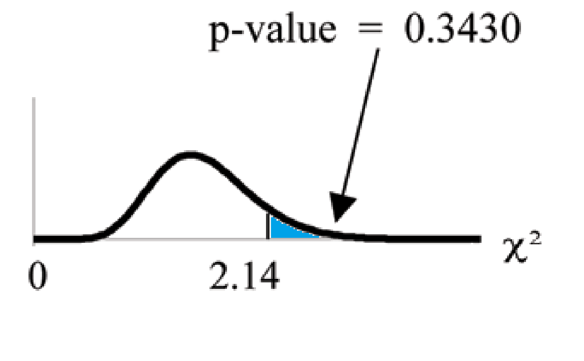
H_o: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where df $= 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



Probability statement: p-value $= P(\chi^2 > 2.14) = 0.3430$

Compare α and the p-value:

- $\alpha = 0.05$
- p-value $= 0.3430$

So, $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject H_o.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Note:TI-83+ and some TI- 84 calculators: Press

`STAT`

and

`ENTER`

. Make sure you clear lists

`L1`

,

`L2`

, and

`L3`

if they have data in them. Into

L1

, put the observed frequencies

20

,

57

,

23

. Into

L2

, put the expected frequencies

25

,

50

,

25

. Arrow over to list

L3

and up to the name area

"L3"

. Enter

(L1-L2)^2/L2

and

ENTER

. Press

2nd QUIT

. Press

2nd LIST

and arrow over to

MATH

. Press

5

. You should see

"sum"

.

Enter L3

. Rounded to 2 decimal places, you should see

2.14

. Press

2nd DISTR

. Arrow down to

7:χ2cdf

(or press

7

). Press

ENTER

. Enter

2.14,1E99,2)

. Rounded to 4 places, you should see

`.3430`

which is the p-value.

The newer TI-84 calculators have in

`STAT TESTS`

the test

`Chi2 GOF`

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

`STAT`

`TESTS`

and

`Chi2 GOF`

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

`calculate`

or

`draw`

. Make sure you clear any lists before you start.

Test of Independence

This module describes how the chi-square distribution can be used to test for independence.

Tests of independence involve using a **contingency table** of observed (data) values. You first saw a contingency table when you studied probability in the Probability Topics chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

Equation:

$$\sum_{(i\cdot j)} \frac{(O-E)^2}{E}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

Note: The expected value for each cell needs to be at least 5 in order to use this test.

Example:

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A \text{ AND } B) = P(A)P(B)$. A AND B is the event that a driver received a speeding violation last year and is also a cell phone user while driving. Suppose, in a study of drivers who received speeding violations in the last year and who uses cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were cell phone users while driving and 450 were not.

Let y = expected number of drivers that use a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$

Solve for $y: y = \frac{70 \cdot 305}{755} = 28.3$

About 28 people from the sample are expected to be cell phone users while driving and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**.

If we do a test of independence using the example above, then the null hypothesis is:

H_o: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be cell phone users while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

The degrees of freedom for the test of independence are:

$$\text{df} = (\text{number of columns - }1)(\text{number of rows - }1)$$

The following formula calculates the **expected number** (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

Example:

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Type of Volunteer	**1-3 Hours**	**4-6 Hours**	**7-9 Hours**	**Row Total**
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Number of Hours Worked Per Week by Volunteer Type (Observed)The table contains **observed (O)** values (data).

Exercise:

Problem: Are the number of hours volunteered **independent** of the type of volunteer?

Solution:

The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_o: The number of hours volunteered is **independent** of the type of volunteer.

H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected table is:

Type of Volunteer	**1-3 Hours**	**4-6 Hours**	**7-9 Hours**
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Number of Hours Worked Per Week by Volunteer Type (Expected)The table contains **expected** (E) values (data).

For example, the calculation for the expected frequency for the top left cell is

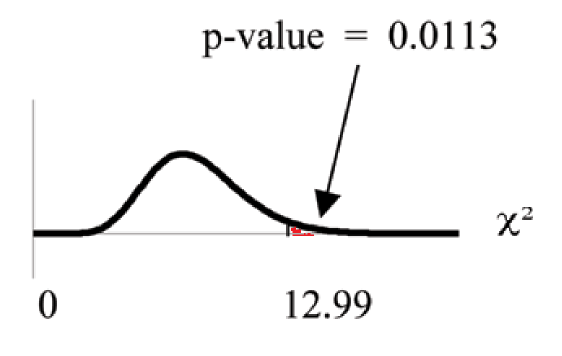
$$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}} = \frac{255 \cdot 298}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$\text{df} = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$

Graph:



Probability statement: p-value $= P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$.
p-value $= 0.0113$. $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Note: Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Example 11-6. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the p-value $= 0.0113$. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Example:
De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School	**High Anxiety**	**Med-high Anxiety**	**Medium Anxiety**	**Med-low Anxiety**	**Low Anxiety**	**Row Total**
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Need to Succeed in School vs. Anxiety Level

Exercise:

Problem:

How many high anxiety level students are expected to have a high need to succeed in school?

Solution:

The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Exercise:

Problem:

If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution:

The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Exercise:

Problem:

- **a** $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} =$
- **b** The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

Solution:

- **a** $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} = 8.19$
- **b** 8

Glossary

Contingency Table

The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

Summary of Formulas

This module provides a summary on formulas used in Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The Chi-Square Probability Distribution

$\mu = \text{df}$ and $\sigma = \sqrt{2 \cdot \text{df}}$

Goodness-of-Fit Hypothesis Test

- Use goodness-of-fit to test whether a data set fits a particular probability distribution.
- The degrees of freedom are $\text{number of cells or categories - 1}$.
- The test statistic is $\sum\limits_{k} \frac{(O-E)^2}{E}$, where O = observed values (data), E = expected values (from theory), and k = the number of different data cells or categories.
- The test is right-tailed.

Test of Independence

- Use the test of independence to test whether two factors are independent or not.
- The degrees of freedom are equal to $(\text{number of columns - 1})(\text{number of rows - 1})$.
- The test statistic is $\sum\limits_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Test of Homogeneity

- Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other.
- The degrees of freedom are equal to $\text{number of columns - 1}$.

- The test statistic is $\sum\limits_{(i\cdot j)}\frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

Note: The expected value for each cell needs to be at least 5 in order to use the Goodness-of-Fit, Independence and Homogeneity tests.

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom are the number of samples - 1.
- The test statistic is $\frac{n-1\ \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left, right, or two-tailed.

Practice 1: Goodness-of-Fit Test
This module provides a practice on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a goodness-of-fit test.

Given

The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as follows: (Source: *HIV/AIDS Epidemiology Santa Clara County, Santa Clara County Public Health Department, May 2011*)

Ethnicity	**Number of Cases**
White	2229
Hispanic	1157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4075

The percentage of each ethnic group in Santa Clara County is as follows:

Ethnicity	**Percentage of total county population**	**Number expected (round to 2 decimal places)**
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Expected Results

If the ethnicity of AIDS victims followed the ethnicity of the total county population, fill in the expected number of cases per ethnic group.

Goodness-of-Fit Test

Perform a goodness-of-fit test to determine whether the make-up of AIDS cases follows the ethnicity of the general population of Santa Clara County.

Exercise:

Problem: H_o:

Exercise:

Problem: H_a:

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Exercise:

Problem: degrees of freedom =

Solution:

degrees of freedom = 3

Exercise:

Problem: test statistic =

Solution:

2016.14

Exercise:

Problem: p-value =

Solution:

Rounded to 4 decimal places, the p-value is 0.0000.

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Let α

Decision:

Reason for the Decision:

Conclusion (write out in complete sentences):

Discussion Question

Exercise:

Problem:

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

Practice 2: Contingency Tables
This module provides a practice on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a test for independence using contingency tables.

Conduct a hypothesis test to determine if smoking level and ethnicity are independent.

Collect the Data

Copy the data provided in **Probability Topics Practice 1: Contingency Tables** into the table below.

Smoking Level Per Day	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
1-10						
11-20						
21-30						
31+						
TOTALS						

Smoking Levels by Ethnicity (Observed)

Hypothesis

State the hypotheses.

- H_o:
- H_a:

Expected Values

Enter expected values in the above below. Round to two decimal places.

Analyze the Data

Calculate the following values:

Exercise:

Problem: Degrees of freedom =

Solution:

12

Exercise:

Problem: test statistic =

Solution:

10301.8

Exercise:

Problem: p-value =

Solution:

0

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Solution:

right

Graph the Data

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Conclusions

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α .

Exercise:

Problem: α

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Solution:

- **a**Reject the null hypothesis

Exercise:

Problem: α

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Homework

This module provides homework on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Exercise:

Problem:

- **a**Explain why the “goodness of fit” test and the “test for independence” are generally right tailed tests.
- **b**If you did a left-tailed test, what would you be testing?

Word Problems

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to The Table of Contents 14. Appendix for the chi-square solution sheet. Round expected frequency to two decimal places.

Exercise:

Problem:

A 6-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data below are the result of the 120 rolls.

Face Value	**Frequency**	**Expected Frequency**
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

Exercise:

Problem:

The marital status distribution of the U.S. male population, age 15 and older, is as shown below. (*Source: U.S. Census Bureau, Current Population Reports*)

Marital Status	Percent	Expected Frequency
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 – 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in the above table, rounding to two decimal places.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

Solution:

- **a**The data fits the distribution

- **b**The data does not fit the distribution
- **c**3
- **e**19.27
- **f**0.0002
- **h**Decision: Reject Null; Conclusion: Data does not fit the distribution.

The next two questions refer to the following information. The columns in the chart below contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class and the Overall Student Population. (*Source: http://www.collegeboard.com*). Suppose the right column contains the result of a survey of 1000 local students from that year who took an AP Exam.

Race/Ethnicity	**AP Examinee Population**	**Overall Student Population**	**Survey Frequency**
Asian, Asian American or Pacific Islander	10.2%	5.4%	113
Black or African American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

Exercise:

Problem:

Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U. S. Overall Student Population based on ethnicity.

Exercise:

Problem:

Perform a goodness-of-fit test to determine whether the local results follow the distribution of U. S. AP Examinee Population, based on ethnicity.

Solution:

- **c**5
- **e**13.4
- **f**0.0199
- **g**Decision: Reject null when $\alpha = 0.05$; Conclusion: Local data do not fit the AP Examinee Distribution. Decision: Do not reject null when $\alpha = 0.01$; Conclusion: There is insufficient evidence to conclude that Local data do not fit the AP Examinee Distribution.

Exercise:

Problem:

The City of South Lake Tahoe, CA, has an Asian population of 1419 people, out of a total population of 23,609 (*Source: U.S. Census Bureau*). Suppose that a survey of 1419 self-reported Asians in Manhattan, NY, area yielded the data in the table below. Conduct a goodness of fit test to determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Race	**Lake Tahoe Frequency**	**Manhattan Frequency**	
Asian Indian	131	174	
Chinese	118	557	
Filipino	1045	518	
Japanese	80	54	
Korean	12	29	
Vietnamese	9	21	
Other	24	66	

The next two questions refer to the following information: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of student expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5000 graduating females and 5000 graduating males was done as a follow-up last year to determine what their actual major was. The results are shown in the tables for Exercises 7 and 8. The second column in each table does not add to 100% because of rounding.

Exercise:

Problem:

Conduct a hypothesis test to determine if the actual college major of graduating females fits the distribution of their expected majors.

Major	Women - Expected Major	Women - Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

Solution:

- **c**10

- **e**11.48
- **f**0.3214
- **h**Decision: Do not reject null when $a = 0.05$ and $a = 0.01$; Conclusion: There is insufficient evidence to conclude that the distribution of majors by graduating females does not fit the distribution of expected majors.

Exercise:

Problem:

Conduct a hypothesis test to determine if the actual college major of graduating males fits the distribution of their expected majors.

Major	**Men - Expected Major**	**Men - Actual Major**
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

Exercise:

Problem:

A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Solution:

- **c**4
- **e**10.53
- **f**0.0324
- **h**Decision: Reject null; Conclusion: Best ski area and level of skier are not independent.

Exercise:

Problem:

Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the following results. Conduct a test for independence.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
2	20	50	70	80
3 - 4	20	50	100	90
5+	20	30	70	70

Exercise:

Problem:

College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Below are the data. Conduct a test for independence.

Major	< $50,000	$50,000 - $68,999	$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Solution:

- **c**8
- **e**33.55
- **f**0
- **h**Decision: Reject null; Conclusion: Major and starting salary are not independent events.

Exercise:

Problem:

Some travel agents claim that honeymoon hot spots vary according to age of the bride and groom. Suppose that 280 East Coast recent brides were interviewed as to where they spent their honeymoons. The information is given below. Conduct a test for independence.

Location	**20 - 29**	**30 - 39**	**40 - 49**	**50 and over**
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Exercise:

Problem:

A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test for independence.

Sport	**18 - 25**	**26 - 30**	**31 - 40**	**41 and over**
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

Solution:

- **c**6
- **e**25.21
- **f**0.0003
- **h**Decision: Reject null

Exercise:

Problem:

A major food manufacturer is concerned that the sales for its skinny French fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are below. Conduct a test for independence.

Type of Fries	Northeast	South	Central	West
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

Exercise:

Problem:

According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $200,000	$200,000 - $400,000	$401,001 - $1,000,000	$1,000,000 +
20 - 29	40	15	40	0	5

Age of Males	None	< $200,000	$200,000 - $400,000	$401,001 - $1,000,000	$1,000,000 +
30 - 39	35	5	20	20	10
40 - 49	20	0	30	0	30
50 +	40	30	15	15	10

Solution:

- **c**12
- **e**125.74
- **f**0
- **h**Decision: Reject null

Exercise:

Problem:

Suppose that 600 thirty–year–olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test for independence.

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< $30,000	15	25	10	5
$30,000 - $40,000	20	40	70	30
$40,000 - $50,000	10	20	40	55
$50,000 - $60,000	5	10	20	60

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
$60,000 +	0	5	10	150

Exercise:

Problem:

A Psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown below. Conduct a Test of Homogeneity. Test at a 5% level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

Solution:

- **c**4
- **d**Chi-Square with df = 4
- **e**3.01
- **f**p-value = 0.5568
- **h**
  ii. Do not reject the null hypothesis.
  iv. There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors.

Exercise:

Problem:

Do men and women select different breakfasts? The breakfast ordered by randomly selected men and women at a popular breakfast place is shown below. Conduct a test of homogeneity. Test at a 5% level of significance

	French Toast	Pancakes	Waffles	Omelettes
Men	47	35	28	53
Women	65	59	55	60

Solution:

- **c**3
- **e**4.01
- **f**p-value = 0.2601
- **h**
  ii. Do not reject the null hypothesis.
  iv. There is insufficient evidence to conclude that the distribution of breakfast ordered is different for men and women.

Exercise:

Problem:

Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of the randomly selected community college students 43 used a computer, 102 used a calculator with built in statistics functions, and 65 used a table from the textbook. Of the randomly selected university students 28 used a computer, 33 used a calculator with built in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.

Solution:

- **c**2
- **e**7.05
- **f**p-value = 0.0294
- **h**
  ii. Reject the null hypothesis.
  iv. There is sufficient evidence to conclude that the distribution of technology use for statistics homework is not the same for statistics students at community colleges and at universities.

Exercise:

Problem:

A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform the hypothesis test at a 5% level of significance.

Solution:

- **c**3
- **d**Chi-Square with df = 3
- **e**11.75
- **f**p-value = 0.0083
- **h**
  ii. Reject the null hypothesis.
  iv. There is sufficient evidence to conclude that the distribution of fish in Green Valley Lake is not the same as the distribution of fish in Echo Lake.

Exercise:

Problem:

A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most $\frac{1}{2}$ oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

Solution:

- **c**83
- **d**Chi-Square with df = 83
- **e**96.81
- **f**p-value = 0.1426; There is a 0.1426 probability that the sample standard deviation is 0.54 or more.
- **h**Decision: Do not reject null; Conclusion: There is insufficient evidence to conclude that the standard deviation is more than 0.5 oz. It cannot be determined whether the equipment needs to be recalibrated or not.

Exercise:

Problem:

Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

Exercise:

Problem:

Isabella, an accomplished **Bay to Breakers** runner, claims that the standard deviation for her time to run the 7 ½ mile race is at most 3 minutes. To test her claim, Rupinder looks up 5 of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

Solution:

- **c**4
- **d**Chi-Square with df = 4
- **e**4.52
- **f**0.3402
- **h**Decision: Do not reject null.

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is 6 with a variance of 9 at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

Exercise:

Problem:

The number of births per woman in China is 1.6 down from 5.91 in 1966 (Source*World Bank, 6/5/12*). This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had. Below are the results. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of births	Frequency
0	5
1	30
2	10
3	5

Solution:

- **c**49
- **d**Chi-Square with df = 49
- **e**54.37
- **f**p-value = 0.2774; If the null hypothesis is true, there is a 0.2774 probability that the sample standard deviation is 0.79 or more.
- **h**Decision: Do not reject null; Conclusion: There is insufficient evidence to conclude that the standard deviation is more than 0.75. It cannot be determined if the standard deviation is greater than 0.75 or not.

Exercise:

Problem:

According to an avid aquariest, the average number of fish in a 20–gallon tank is 10, with a standard deviation of 2. His friend, also an aquariest, does not believe that the standard deviation is 2. She counts the number of fish in 15 other 20–gallon tanks. Based on the results that follow, do you think that the standard deviation is different from 2? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

Exercise:

Problem:

The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a 10–ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of 2 oz.

Solution:

- **a** $\sigma^2 \leq (1.5)^2$
- **c**48
- **d**Chi-Square with df = 48

- **e**85.33
- **f**0.0007
- **h**Decision: Reject null.

Try these true/false questions.

Exercise:

Problem:

As the degrees of freedom increase, the graph of the chi-square distribution looks more and more symmetrical.

Solution:

True

Exercise:

Problem: The standard deviation of the chi-square distribution is twice the mean.

Solution:

False

Exercise:

Problem:

The mean and the median of the chi-square distribution are the same if $\text{df} = 24$.

Solution:

False

Exercise:

Problem:

In a Goodness-of-Fit test, the expected values are the values we would expect if the null hypothesis were true.

Solution:

True

Exercise:

Problem:

In general, if the observed values and expected values of a Goodness-of-Fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

Solution:

True

Exercise:

Problem:

The degrees of freedom for a Test for Independence are equal to the sample size minus 1.

Solution:

False

Exercise:

Problem:

Use a Goodness-of-Fit test to determine if high school principals believe that students are absent equally during the week or not.

Solution:

True

Exercise:

Problem: The Test for Independence uses tables of observed and expected data values.

Solution:

True

Exercise:

Problem:

The test to use when determining if the college or university a student chooses to attend is related to his/her socioeconomic status is a Test for Independence.

Solution:

True

Exercise:

Problem: The test to use to determine if a six-sided die is fair is a Goodness-of-Fit test.

Solution:

True

Exercise:

Problem:

In a Test of Independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

Solution:

True

Exercise:

Problem:

In a Goodness-of Fit test, if the p-value is 0.0113, in general, do not reject the null hypothesis.

Solution:

False

Exercise:

Problem:

For a Chi-Square distribution with degrees of freedom of 17, the probability that a value is greater than 20 is 0.7258.

Solution:

False

Exercise:

Problem:

If $\text{df} = 2$, the chi-square distribution has a shape that reminds us of the exponential.

Solution:

True

Review

This module provides an review on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The next two questions refer to the following real study:

A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12 - 19. 6% of the girls surveyed said they have been pregnant. (*Parade Magazine*) We are interested in the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Exercise:

Problem:

Find the 95% confidence interval for the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Solution:

$(0.0424,0.0770)$

Exercise:

Problem:

The report also stated that the results of the survey are accurate to within ± 3.7% at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

Solution:

2401

Exercise:

Problem:

Given: X ~ $\text{Exp}\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

The next four questions refer to the following information:

Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of 7 years and a standard deviation of 2 years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

Exercise:

Problem:

60% of individuals keep their cars **at most** how many years?

Solution:

7.5

Exercise:

Problem:

Suppose that we randomly survey one person. Find the probability that person keeps his/her car **less than** 2.5 years.

Solution:

0.0122

Exercise:

Problem:

If we are to pick individuals 10 at a time, find the distribution for the **mean** car length ownership.

Solution:

$N(7,0.63)$

Exercise:

Problem:

If we are to pick 10 individuals, find the probability that the **sum** of their ownership time is more than 55 years.

Solution:

0.9911

Exercise:

Problem: For which distribution is the median not equal to the mean?

- **A**Uniform
- **B**Exponential
- **C**Normal
- **D**Student-t

Solution:

B

Exercise:

Problem:

Compare the standard normal distribution to the student-t distribution, centered at 0. Explain which of the following are true and which are false.

- **a**As the number surveyed increases, the area to the left of -1 for the student-t distribution approaches the area for the standard normal distribution.
- **b**As the degrees of freedom decrease, the graph of the student-t distribution looks more like the graph of the standard normal distribution.
- **c**If the number surveyed is 15, the normal distribution should never be used.

Solution:

- **a**True
- **b**False
- **c**False

The next five questions refer to the following information:

We are interested in the checking account balance of a twenty-year-old college student. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of $640 and a sample standard deviation of $150. Let X = checking account balance of an individual twenty year old college student.

Exercise:

Problem: Explain why we cannot determine the distribution of X.

Exercise:

Problem:

If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of 20-year old college students, what distribution would you use?

Solution:

student-t with df $= 15$

Exercise:

Problem:

Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

Solution:

$(560.07,719.93)$

Exercise:

Problem:

What type of data is the balance of the checking account considered to be?

Solution:

quantitative - continuous

Exercise:

Problem:

What type of data is the number of 20 year olds considered to be?

Solution:

quantitative - discrete

Exercise:

Problem:

On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

- **a**Define the random variable X.
- **b**State the distribution for X.
- **c**Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Solution:

- **b** $P(4)$
- **c**0.0183

The next two questions refer to the following information:

The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30 . Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

Exercise:

Problem:

Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

Solution:

greater than

Exercise:

Problem:

Is it likely that exactly 8 of the 15 plays would pay back money? Justify your answer numerically.

Solution:

No; $P(x = 8) = 0.0348$

Exercise:

Problem: A game is played with the following rules:

- it costs $10 to enter
- a fair coin is tossed 4 times
- if you do not get 4 heads or 4 tails, you lose your $10
- if you get 4 heads or 4 tails, you get back your $10, plus $30 more

Over the long run of playing this game, what are your expected earnings?

Solution:

You will lose $5

Exercise:

Problem:

- The mean grade on a math exam in Rachel’s class was 74, with a standard deviation of 5. Rachel earned an 80.
- The mean grade on a math exam in Becca’s class was 47, with a standard deviation of 2. Becca earned a 51.
- The mean grade on a math exam in Matt’s class was 70, with a standard deviation of 8. Matt earned an 83.

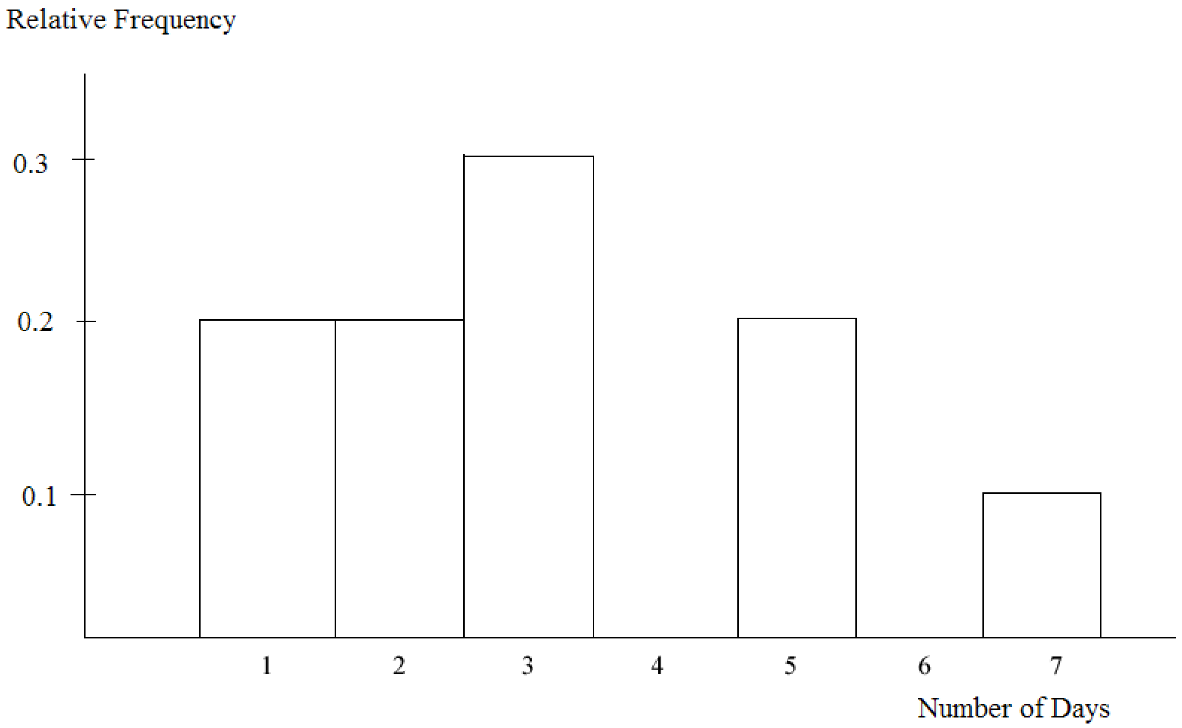
Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Solution:

Becca

The next two questions refer to the following information:

A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:



Exercise:

Problem: Find the number of responses that were “5".

Solution:

14

Exercise:

Problem:

Find the mean, standard deviation, the median, the first quartile, the third quartile and the IQR.

Solution:

- Sample mean = 3.2
- Sample standard deviation = 1.85
- Median = 3
- Quartile 1 = 2

- Quartile 3 = 5
- IQR = 3

Exercise:

Problem:

Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home.

Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

- **a**State an appropriate **null** hypothesis.
- **b**State an appropriate **alternate** hypothesis.
- **c**Define the Random Variable, P'.
- **d**Calculate the test statistic.
- **e**Calculate the p-value.
- **f**At the 5% level of decision, what is your decision about the null hypothesis?
- **g**What is the Type I error?
- **h**What is the Type II error?

Solution:

- **d** $z = -1.19$
- **e**0.1171
- **f**Do not reject the null

Exercise:

Problem:

Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

Solution:

We conclude that the patient does have the HIV virus when, in fact, the patient does not.

Exercise:

Problem:

It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. 15 Californians wear jeans to work and 6 non-Californians wear jeans to work.

- C = Californian professional
- NC = non-Californian professional

- **a**State appropriate **null** and **alternate** hypotheses.
- **b**Define the Random Variable.
- **c**Calculate the test statistic and p-value.
- **d**At the 5% significance level, what is your decision?
- **e**What is the Type I error?
- **f**What is the Type II error?

Solution:

- **c** $z = 2.21$; $p = 0.0136$
- **d**Reject the null

- **e**We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
- **f**We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

The next two questions refer to the following information:

A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

Exercise:

Problem: This is a test of (pick the best answer):

- **A**large samples, independent means
- **B**small samples, independent means
- **C**dependent means

Solution:

C

Exercise:

Problem: State the distribution to use for the test.

Solution:

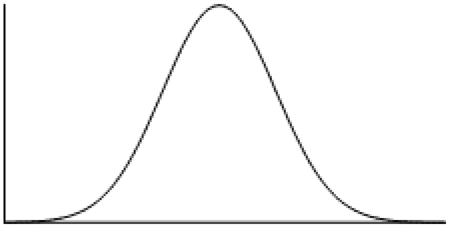
t_5

Solution Sheet: Hypothesis Testing for Single Mean and Single Proportion
This module provides a solution sheet for the Hypothesis Testing: Single Mean and Single Proportion chapter of the Collaborative Statistics textbook/collection.

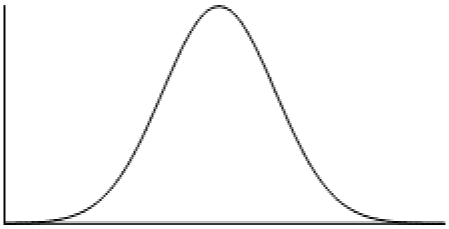
Class Time:

Name:

- **a**H_o:
- **b**H_a:
- **c**In words, **CLEARLY** state what your random variable X or P represents.
- **d**State the distribution to use for the test.
- **e**What is the test statistic?
- **f**What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.
- **g**Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



- **h**Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
    - **i**Alpha:
    - **ii**Decision:
    - **iii**Reason for decision:
    - **iv**Conclusion:
- **i**Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.



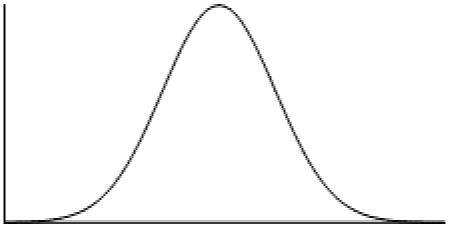
Solution Sheet: Hypothesis Testing for Two Means, Paired Data, and Two Proportions

This module provides a solution sheet for the Hypothesis Testing: Two Means, Paired Data, Two Proportions chapter of the Collaborative Statistics textbook/collection.

Class Time:

Name:

- **a**H_o: ________
- **b**H_a: ________
- **c**In words, **clearly** state what your random variable X X , P P - or X_d represents.
- **d**State the distribution to use for the test.
- **e**What is the test statistic?
- **f**What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.
- **g**Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



- **h**Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
  - **i**Alpha:
  - **ii**Decision:
  - **iii**Reason for decision:
  - **iv**Conclusion:
- **i**In complete sentences, explain how you determined which distribution to use.

Solution Sheet: The Chi-Square Distribution

This module provides a solution sheet for the Chi-Square Distribution chapter of the Collaborative Statistics textbook/collection.

Class Time:

Name:

- **a**H_o: ________
- **b**H_a: ________
- **c**What are the degrees of freedom?
- **d**State the distribution to use for the test.
- **e**What is the test statistic?
- **f**What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.
- **g**Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

- **h**Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences.**
    - **i**Alpha:
    - **ii**Decision:
    - **iii**Reason for decision:
    - **iv**Conclusion:

Symbols and their Meanings
This module defines symbols used throughout the Collaborative Statistics textbook.

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Sampling and Data	√	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q1	Quartile one	the first quartile
Descriptive Statistics	Q2	Quartile two	the second quartile
Descriptive Statistics	Q3	Quartile three	the third quartile
Descriptive Statistics	IQR	inter-quartile range	Q3-Q1=IQR
Descriptive Statistics	x	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Descriptive Statistics	s s_x sx	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s-squared	sample variance
Descriptive Statistics	σ σ_x σx	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma-squared	population variance
Descriptive Statistics	$\mathit{\Sigma}$	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Probability Topics	$P\left(A \mid B\right)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P\left(A\text{ or }B\right)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P\left(A\text{ and }B\right)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A’	A-prime, complement of A	complement of A, not A
Probability Topics	$P\left(\text{A'}\right)$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P\left(G_1\right)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	X~	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	$\neq$	not equal to	same
Continuous Random Variables	$f\left(x\right)$	f of x	function of x
Continuous Random Variables	pdf	prob. density function	same
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	Exp	exponential distribution	same
Continuous Random Variables	k	k	critical value

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	X	X-bar	the random variable X-bar
The Central Limit Theorem	μ_x	mean of X	the average of X
The Central Limit Theorem	μ_x	mean of X-bar	the average of X-bar

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
The Central Limit Theorem	σ_x	standard deviation of X	same
The Central Limit Theorem	σ_x	standard deviation of X-bar	same
The Central Limit Theorem	ΣX	sum of X	same
The Central Limit Theorem	Σx	sum of x	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same
Confidence Intervals	t	student-t distribution	same
Confidence Intervals	df	degrees of freedom	same

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student-t with a/2 area in right tail	same
Confidence Intervals	p’ $\widehat{p}$	p-prime; p-hat	sample proportion of success
Confidence Intervals	q’ $\hat{q}$	q-prime; q-hat	sample proportion of failure
Hypothesis Testing	H_0	H-naught, H-sub 0	null hypothesis
Hypothesis Testing	H_a	H-a, H-sub a	alternate hypothesis
Hypothesis Testing	H_1	H-1, H-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$X1 - X2$	X1-bar minus X2-bar	difference in sample means

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
	$\mu_1 - \mu_2$	mu-1 minus mu-2	difference in population means
	$P'_1 - P'_2$	P1-prime minus P2-prime	difference in sample proportions
	$p_1 - p_2$	p1 minus p2	difference in population proportions
Chi-Square Distribution	X^2	Ky-square	Chi-square
	O	Observed	Observed frequency
	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + \text{bx}$	y equals a plus b-x	equation of a line
	$\hat{y}$	y-hat	estimated value of y
	r	correlation coefficient	same

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
	ε	error	same
	SSE	Sum of Squared Errors	same
	$1.9s$	1.9 times s	cut-off value for outliers
F-Distribution and ANOVA	F	F-ratio	F ratio

Symbols and their Meanings

Formulas

This module provides an overview of Statistics Formulas used as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Formula

Factorial

$$n! = n(n-1)(n-2)\ldots(1)$$

$$0! = 1$$

Formula

Combinations

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Formula

Binomial Distribution

$$X\text{~}B(n,p)$$

$P\big(X=x\big) = \binom{n}{x}p^x q^{n-x}$, for $x = 0, 1, 2, \ldots, n$

Formula

Geometric Distribution

$$X\text{~}G(p)$$

$P\big(X=x\big) = q^{x-1}p$, for $x = 1, 2, 3, \ldots$

Formula

Hypergeometric Distribution

X~H (r, b, n)

$$P\left(X=x\right) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}}\right)$$

Formula

Poisson Distribution

X~P (μ)

$$P\left(X = x\right) = \frac{\mu^{x}e^{-\mu}}{x!}$$

Formula

Uniform Distribution

$$X\text{~}U(a,b)$$

$$f\left(X\right) = \frac{1}{b - a}, a < x < b$$

Formula

Exponential Distribution

$$X\text{~ Exp}(m)$$

$$f(x) = me^{-\text{mx}}, m > 0, x \geq 0$$

Formula

Normal Distribution

$$X\text{~}N\left(\mu,\sigma^{2}\right)$$

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x - \mu)^{2}}{2\sigma^{2}}}\ , \qquad -\infty < x < \infty$$

Formula

Gamma Function

$$\mathit{\Gamma}\left(z\right) = \int_{0}^{\infty}x^{z - 1}e^{-x}\,\text{dx}\ z > \mathit{0}$$

$$\mathit{\Gamma}\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$\mathit{\Gamma}(m + 1) = m!$ for m, a nonnegative integer

otherwise: $\mathit{\Gamma}(a + 1) = \text{a}\Gamma(a)$

Formula

Student-t Distribution

$$X\text{~}t_{\text{df}}$$

$$f\left(x\right)=\frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}}\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\mathrm{n\pi}}\Gamma\left(\frac{n}{2}\right)}$$

$$X=\frac{Z}{\sqrt{\frac{Y}{n}}}$$

Z~$N(0,1)$, Y~X_{df}^2 ,n = degrees of freedom

Formula

Chi-Square Distribution

X~X_{df}^2

$f\left(x\right)=\frac{x^{\frac{n-2}{2}}e^{\frac{-x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$, $x>0$, n = positive integer and degrees of freedom

Formula

F Distribution

X~$F_{\mathrm{df}(n),\mathrm{df}(d)}$

$\mathrm{df}(n)=$degrees of freedom for the numerator

$\mathrm{df}(d)=$degrees of freedom for the denominator

$$f\left(x\right)=\frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)}\left(\frac{u}{v}\right)^{\frac{u}{2}}x^{\left(\frac{u}{2}-1\right)}\left[1+\left(\frac{u}{v}\right)x^{-.5(u+v)}\right]$$

$X=\frac{Y_u}{W_v}$, Y, W are chi-square

Notes for the TI-83, 83+, 84 Calculator
Notes and tips for using TI-83, TI-83+, and TI-84 calculators for statistics applications.

Quick Tips

Legend

- ▭

  represents a button press
- `[ ]` represents yellow command or green letter behind a key
- `< >` represents items on the screen

To adjust the contrast
Press

2nd

, then hold

▲

to increase the contrast or

▼

to decrease the contrast.

To capitalize letters and words
Press

ALPHA

to get one capital letter, or press

2nd

, then

ALPHA

to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake
If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation
Numbers in scientific notation are expressed on the TI-83, 83+, and 84 using E notation, such that...

- 4.321 E 4 = 4.321×10^4
- 4.321 E -4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another:
Both calculators: Insert your respective end of the link cable cable and press

2nd

, then `[LINK]`.
Calculator receiving information:

Use the arrows to navigate to and select`<RECEIVE>`
Press
ENTER

Calculator sending information:

Press appropriate number or letter.

Use up and down arrows to access the appropriate item.
Press ENTER to select item to transfer.

Press right arrow to navigate to and select`<TRANSMIT>`.
Press ENTER

Note:ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press

2nd

, then `[QUIT]` To exit when done.

Manipulating One-Variable Statistics

Note: These directions are for entering data with the built-in statistical program.

Data	Frequency
-2	10

Data	Frequency
-1	3
0	4
1	5
3	8

Sample DataWe are manipulating 1-variable statistics.

To begin:

Turn on the calculator.

ON

Access statistics mode.

STAT

Select`<4:ClrList>`to clear data from lists, if desired.

4

,

ENTER

Enter list`[L1]`to be cleared.

2nd

, `[L1]` ,

ENTER

Display last instruction.

2nd

, [ENTRY]

Continue clearing remaining lists in the same fashion, if desired.

◀

,

2nd

, [L2] ,

ENTER

Access statistics mode.

STAT

Select <1:Edit . . .>

ENTER

Enter data. Data values go into [L1]. (You may need to arrow over to [L1])

- Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad)

(−)

,

	9 , ENTER - Continue in the same manner until all data values are entered.
In [L2], enter the frequencies for each data value in [L1].	- Type in a frequency and enter it. (If a data value appears only once, the frequency is "1") 4 , ENTER - Continue in the same manner until all data values are entered.
Access statistics mode.	STAT
Navigate to <CALC>	
Access <1:1-var Stats>	ENTER
Indicate that the data is in [L1]...	2nd , [L1] ,

,

...and indicate that the frequencies are in`[L2]`.

2nd

, `[L2]` ,

ENTER

The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

Note: We will assume that the data is already entered

We will construct 2 histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

Access graphing mode.

2nd

, `[STAT PLOT]`

Select`<1:plot 1>`To access plotting - first graph.

ENTER

Use the arrows navigate go to`<ON>`to turn on Plot 1.

`<ON>` .

,

ENTER

Use the arrows to go to the histogram picture and select the histogram. ENTER

Use the arrows to navigate to<Xlist>
If "L1" is not selected, select it.

2nd

, [L1] ,

ENTER

Use the arrows to navigate to<Freq>.
Assign the frequencies to[L2].

2nd

, [L2] ,

ENTER

Go back to access other graphs.

2nd

, [STAT PLOT]
Use the arrows to turn off the remaining plots.
Be sure to deselect or clear all equations before graphing.

To deselect equations:

Access the list of equations.

Y=

Select each equal sign (=).

▼

▶

ENTER

Continue, until all equations are deselected.

To clear equations:

Access the list of equations.

Y=

Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

▼

▶

CLEAR

Repeat until all equations are deleted.

To draw default histogram:

Access the ZOOM menu.

ZOOM

Select`<9:ZoomStat>`

9

The histogram will show with a window automatically set.

To draw custom histogram:

Access WINDOW to set the graph parameters.

- $X_{\text{min}} = -2.5$
- $X_{\text{max}} = 3.5$
- $X_{\text{scl}} = 1$ (width of bars)
- $Y_{\text{min}} = 0$
- $Y_{\text{max}} = 10$
- $Y_{\text{scl}} = 1$ (spacing of tick marks on y-axis)
- $X_{\text{res}} = 1$

Access GRAPH to see the histogram.

To draw box plots:

Access graphing mode.

2nd

, `[STAT PLOT]`

Select `<1:Plot 1>` to access the first graph.

ENTER

Use the arrows to select `<ON>` and turn on Plot 1.

ENTER

Use the arrows to select the box plot picture and enable it.

ENTER

Use the arrows to navigate to<Xlist>
If "L1" is not selected, select it.

2nd

, [L1] ,

ENTER

Use the arrows to navigate to<Freq>.
Indicate that the frequencies are in[L2].

2nd

, [L2] ,

ENTER

Go back to access other graphs.

2nd

, [STAT PLOT]

Be sure to deselect or clear all equations before graphing using the method mentioned above.
View the box plot.

GRAPH

, [STAT PLOT]

Linear Regression

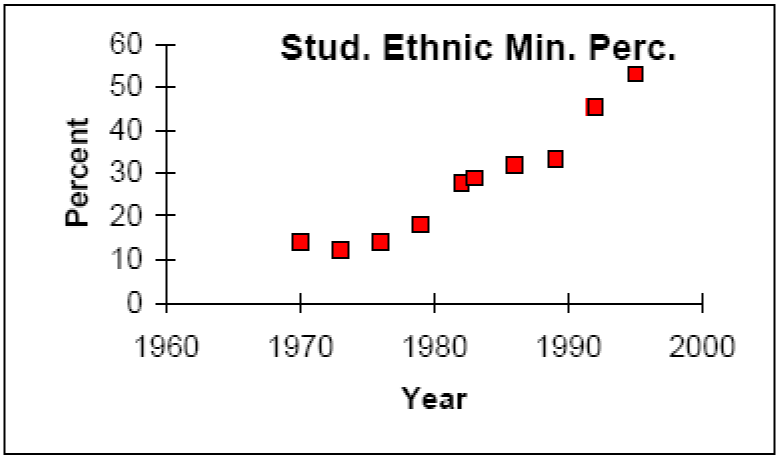
Sample Data

The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970 - 1995 was:

Year	**Student Ethnic Minority Percentage**
1970	14.13
1973	12.27
1976	14.08
1979	18.16
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note: The TI-83 has a built-in linear regression feature, which allows the data to be edited.The x-values will be in

```
[L1]
```

; the y-values in

```
[L2]
```

.

To enter data and do linear regression:

ON Turns calculator on

ON

Before accessing this program, be sure to turn off all plots.

- Access graphing mode

mode.

2nd

, [STAT PLOT]

- Turn off all plots.

4

,

ENTER

Round to 3 decimal places. To do so:

- Access the mode menu.

MODE

, [STAT PLOT]

- Navigate to <Float> and then to the right to <3>.

▼

▶

- All numbers will be rounded to 3 decimal places until changed.

ENTER

Enter statistics mode and clear lists [L1] and [L2], as describe above.

STAT , 4

Enter editing mode to insert values for x and y.

STAT , ENTER

Enter each value. Press ENTER to continue.

To display the correlation coefficient:

Access the catalog.

2nd , [CATALOG]

Arrow down and select <DiagnosticOn>

▼ … , ENTER , ENTER

$r$andr^2will be displayed during regression calculations.
Access linear regression.

STAT

▶

Select the form of$y = a + bx$

8

,

ENTER

The display will show:
LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r^2 = 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- $\text{Percent} = -3176.909 + 1.617(\text{year }\#)$

The correlation coefficient $r = 0.961$
To see the scatter plot:

Access graphing mode.

2nd

, `[STAT PLOT]`

Select `<1:plot 1>` To access plotting - first graph.

ENTER

Navigate and select `<ON>` to turn on Plot 1.

`<ON>`

ENTER

Navigate to the first picture.
Select the scatter plot.

ENTER

Navigate to `<Xlist>`
If `[L1]` is not selected, press 2nd , `[L1]` to select it.

Confirm that the data values are in `[L1]`.

`<ON>`

ENTER

Navigate to `<Ylist>`
Select that the frequencies are in `[L2]`.

2nd

, `[L2]` ,

ENTER

Go back to access other graphs.

2nd

`[STAT PLOT]`

, [STAT PLOT]

Use the arrows to turn off the remaining plots.

Access WINDOW to set the graph parameters.

- $X_{\text{min}} = 1970$
- $X_{\text{max}} = 2000$
- $X_{\text{scl}} = 10$ (spacing of tick marks on x-axis)
- $Y_{\text{min}} = -0.05$
- $Y_{\text{max}} = 60$
- $Y_{\text{scl}} = 10$ (spacing of tick marks on y-axis)
- $X_{\text{res}} = 1$

Be sure to deselect or clear all equations before graphing, using the instructions above.

Press GRAPH to see the scatter plot.

To see the regression graph:

Access the equation menu. The regression equation will be put into Y1.

Y=

Access the vars menu and navigate to `<5: Statistics>`

VARS

,

5

Navigate to `<EQ>`.

`<1: RegEQ>` contains the regression equation which will be entered in Y1.

ENTER

Press GRAPH. The regression line will be superimposed over scatter plot.

To see the residuals and use them to calculate the critical point for an outlier:

Access the list. RESID will be an item on the menu. Navigate to it.

2nd , `[LIST]`, `<RESID>`

Confirm twice to view the list of residuals. Use the arrows to select them.

ENTER , ENTER

The critical point for an outlier is: $1.9V\frac{\text{SSE}}{n-2}$ where:

- ○ n = number of pairs of data
- ○ SSE = sum of the squared errors
- ○ $\sum \text{residual}^2$

Store the residuals in `[L3]`.

STO► , 2nd , `[L3]` , ENTER

Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$

2nd

, [L3] ,

x^2

,

÷

,

8

Store this value in [L4].

STO▶

,

2nd

, [L4] ,

ENTER

Calculate the critical value using the equation above.

1

,

.

,

9

,

X

,

2nd

, [V] ,

2nd

, [LIST]

▶

,

▶

,

5

,

2nd

, [L4] ,

)

,

)

,

ENTER

Verify that the calculator displays: 7.642669563. This is the critical value.

Compare the absolute value of each residual value in `[L3]` to 7.64 . If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of y for various x-values:
There are various ways to determine estimates for "y". One way is to substitute values for "x" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84 instructions for distributions and tests

Distributions

Access `DISTR` (for "Distributions").

For technical assistance, visit the Texas Instruments website at http://www.ti.com and enter your calculator model into the "search" box.
Binomial Distribution

- `binompdf(n,p,x)` corresponds to P(X = x)
- `binomcdf(n,p,x)` corresponds to P(X ≤ x)
- To see a list of all probabilities for x: 0, 1, . . . , n, leave off the "`x`" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to P(X = x)
- `poissoncdf(λ,x)` corresponds to P(X ≤ x)

Continuous Distributions (general)

- $-\infty$ uses the value -1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value (only useful to plot the normal curve, in which case "`x`" is the variable)
- `normalcdf(left bound, right bound, μ,σ)` corresponds to P(left bound < X < right bound)
- `normalcdf(left bound, right bound)` corresponds to P(left bound < Z < right bound) - standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: P(X < k) = p
- `invNorm(p)` yields the critical value, k: P(Z < k) = p for the standard normal

Student-t Distribution

- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-t curve, in which case "`x`" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to P(left bound < t < right bound)

Chi-square Distribution

- X^2`pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "`x`" is the variable)
- X^2`cdf(left bound, right bound, df)` corresponds to P(left bound < X^2 < right bound)

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the F curve, in which case "`x`" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to P(left bound < F < right bound)

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the Confidence Intervals and Hypothesis Tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known
- `TInterval` is the confidence interval for mean when σ is unknown; s estimates σ.
- `1-PropZInt` is the confidence interval for proportion

Note: The confidence levels should be given as percents (ex. enter "`95`" or "`.95`" for a 95% confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known
- `T-Test` is the hypothesis test for single mean when σ is unknown; s estimates σ.
- `2-SampZTest` is the hypothesis test for 2 independent means when both σ's are known
- `2-SampTTest` is the hypothesis test for 2 independent means when both σ's are unknown
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for 2 proportions.
- `X`^{2}`-Test` is the hypothesis test for independence.
- `X`^{2}`GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

Note: Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μ∅`" represents the null hypothesis. For a test of a single proportion, "`p∅`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

Tables

Note: When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- Student-t table
- Normal table
- Chi-Square table
- F-table
- All four tables can be accessed by going to http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm

95% Critical Values of the Sample Correlation Coefficient Table

- 95% Critical Values of the Sample Correlation Coefficient

Note: The url for this table is http://cnx.org/content/m17098/latest/